NAUKA – DYDAKTYKA – PRAKTYKA

Piotr Malak

INDEKSOWANIE TREŚCI

WYDAWNICTWO SBP

Polish Librarians Association

SCIENCE - DIDACTICS - PRACTICE

Piotr Malak

CONTENT INDEXING

Comparizon of efficiency of traditional and automatic methods

WYDAWNICTWO SBP

Warsaw
2012

Stowarzyszenie Bibliotekarzy Polskich
NAUKA - DYDAKTYKA - PRAKTYKA

Piotr Malak

INDEKSOWANIE TREŚCI

Porównanie skuteczności metod tradycyjnych i automatycznych

WYDAWNICTWO SBP

Warszawa
2012

Książka wydana przy pomocy finansowej Instytutu Informacji Naukowej i Bibliologii Uniwersytetu Mikołaja Kopernika.

Recenzenci:
prof. dr hab. Wiesław Babik
prof. dr hab. Irena Kamińska-Szmaj
prof. dr hab. Jadwiga Woźniak-Kasperek

Projekt okładki:
Studio Kałamarnica (www.studiokalamarnica.pl)
Redakcja techniczna i korekta:
Justyna Grzymała

ISBN 978-83-61464-42-6

CIP - Biblioteka Narodowa

Malak, Piotr
Indeksowanie treści : porównanie skuteczności metod tradycyjnych i automatycznych / Piotr Malak ; Stowarzyszenie Bibliotekarzy Polskich. - Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich, 2012. - (Nauka, Dydaktyka, Praktyka ; 133)

Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich
00-355 Warszawa, ul. Konopczyńskiego 5/7, tel. (22) 827-52-96
Warszawa 2012 r. Wyd. I. Ark. wyd. 9,5. Ark. druk. 12,5.
Łamanie: Justyna Grzymała
Druk i oprawa: Drukarnia i Introligatornia OPRAWA Sp. z o.o.
ul. Dowborczyków 17, 90-019 Łódź, tel.(42) 676-42-82

Spis treści

Table of contents

Wstęp

Przyrastająca masowo ilość informacji elektronicznej wymusza, zarówno na jej dostawcach, jak i odbiorcach, stosowanie wydajnych i efektywnych metod jej składowania i dostępu. Zarządzanie informacją przechowywaną w systemach komputerowych wymaga coraz częściej wdrożenia rozwiązań odmiennych od tych, które są stosowane w przypadku nośników tradycyjnych (np. w bibliotekach czy archiwach). Potrzeba ta wynika z wielkości zbiorów, tempa rozrostu kolekcji dokumentów cyfrowych oraz formy, w jakiej są one udostępniane. Z jednej strony mamy więc do czynienia z koniecznością indeksowania treści wielu dokumentów równocześnie, z drugiej zaś, dzięki dostępowi do pełnej treści dokumentów tekstowych, można zaoferować użytkownikowi zaawansowane narzędzia wyszukiwawcze i w krótkim czasie dostarczyć informację w wysokim stopniu relewantną do jego potrzeb.

W przypadku dużych repozytoriów elektronicznych nie ma zazwyczaj czasu ani fizycznych możliwości na przeprowadzenie przez człowieka tradycyjnego opracowania rzeczowego. Dlatego też zautomatyzowane systemy wyszukiwawcze, głównie wyszukiwarki internetowe, indeksując pełną treść dokumentów, dostarczają użytkownikom narzędzi wyszukiwania informacji za pomocą słownictwa niekontrolowanego. Zapytania informacyjne w takich systemach mają bardzo często postać zestawów słów kluczowych, natomiast proces przeszukiwania przeprowadzany jest na peł-

nych tekstach dokumentów. Automatyczne indeksowanie derywacyjne[1] pozwala na wyszukiwanie informacji bez restrykcyjnych ograniczeń formalnych znanych z systemów informacyjno-wyszukiwawczych zarządzających dokumentami tradycyjnymi[2].

Parafrazując słowa Jadwigi Woźniak można stwierdzić, że indeksowanie automatyczne jest swego rodzaju realizacją realistycznej (materialistycznej) teorii przedmiotu dokumentu. Wskazanie wyrażenia bądź słowa kluczowego na podstawie zależności kwantytatywnych jest odnotowaniem obiektywnych cech treściowych dokumentu, bez interpretacji ukierunkowanych na treść albo zapytanie przy założeniu, że czytelnik jest w stanie rozróżnić twierdzenia prawdziwe i fałszywe[3]. Dalej autorka dodaje, że indeksowanie w ujęciu modelowym jest procesem składającym się z analizy dokumentów i zawartych w nich informacji, selekcjonowania informacji ważnych z punktu widzenia relewancji danego siw oraz z tłumaczenia uzyskanego obrazu treści dokumentu na język swobodnych słów kluczowych, a dalej – charakterystyki wyszukiwawczej dokumentu, czyli obrazu dokumentu w stosowanym w systemie języku informacyjnym[4].

Indeksowanie automatyczne nie jest tematem nowym dla informacji naukowej, próby wykorzystania komputerów do zautomatyzowania zarządzania dokumentami i ich treścią były podejmowane od dawna. Za Piotrem Gawrysiakiem można podać, że badania nad statystyczną analizą tekstów były prowadzone przez Markowa już w 1913 roku[5]. Pośrednio o kształtowaniu się zainteresowania tematyką indeksowania automatycznego może świadczyć liczba publikacji na ten temat odnotowanych w bazie LISA. Analizę tych wartości zaprezentowano na wykresie 1. oraz w tabeli 1.

---

[1] Indeksowanie derywacyjne – indeksowanie swobodne wykorzystujące wyrażenia występujące w indeksowanych dokumentach. Jest stosowane w zautomatyzowanych systemach informacyjno-wyszukiwawczych operujących pełnymi tekstami dokumentów. Por. hasło *indeksowanie derywacyjne*. W: *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*, pod red. B. Bojar. Warszawa: Wydawnictwo SBP 2002, s. 86.

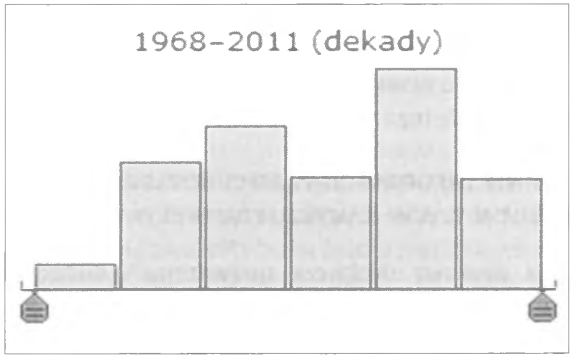
[2] Takimi ograniczeniami mogą być np. znajomość nazw pól z odpowiednimi elementami opisu czy znajomość zastosowanego w danym systemie języka informacyjno-wyszukiwawczego.

[3] J. Woźniak: *Kategoryzacja. Studium z teorii języków informacyjno-wyszukiwawczych*. Warszawa: Wydawnictwo SBP 2000, s. 26.

[4] Tamże, s. 189-190.

[5] Za: P. Gawrysiak: *Klasyfikacja. Narzędzie zarządzania i wyszukiwania informacji*. Warszawa: MOST Press 2009, s. 5.

Wykres 1. Liczba dokumentów poświęconych zagadnieniu indeksowania automatycznego zarejestrowana w bazie LISA.



Źródło: opracowanie własne na podstawie danych z bazy danych LISA: http://search.proquest.com/lisa.

Tabela 1. Liczba dokumentów poświęconych zagadnieniu indeksowania automatycznego zarejestrowana w bazie LISA.

Dekada	Liczba odnotowanych publikacji
1960-1969	55
1970-1979	357
1980-1989	465
1990-1999	316
2000-2009	635
2010-2011	311

Źródło: opracowanie własne na podstawie danych z bazy danych LISA: http://search.proquest.com/lisa.

Niewątpliwie skuteczniejsze i dokładniejsze jest indeksowanie wykonane przez człowieka, jednakże imperatyw czasu dostępu do informacji powoduje, że w systemach indeksujących liczne zbiory informacji stosuje się rozwiązania automatyczne. Z pewnością warto zbadać zagadnienia związane ze skutecznością automatycznego i kognitywnego, będącego bezpośrednim wynikiem procesów poznawczych człowieka, przetwarzania i opracowania informacji. Próbę takiego porównania, w odniesieniu do słów kluczowych jako charakterystyk dokumentów, podjąłem w niniej-

szej książce. Słowa kluczowe stoją w pewnej opozycji do języków stosowanych w tradycyjnych systemach informacyjno-wyszukiwawczych, chociażby przez otwartość, brak kontroli semantycznej słownictwa i maksymalną prostotę reguł składniowych, co powoduje, że nie wymagają od użytkownika specjalistycznego przeszkolenia w wyszukiwaniu. Dlatego też warto rozważyć zalety i wady tego właśnie sposobu reprezentacji treści.

Wyszukiwanie informacji w sieci rozległej: katalogi stron WWW a wyszukiwarki internetowe

Informacja naukowa dostarcza sprawdzonych metod zarządzania treścią dokumentów, przygotowywania charakterystyk wyszukiwawczych oraz sposobów zwiększenia relewancji wyników wyszukiwania w stosunku do zapytania użytkownika. Metody te, w odniesieniu do dokumentów elektronicznych, są z powodzeniem stosowane m.in. w bibliotekach cyfrowych, gdzie zaadaptowano aparat opracowania rzeczowego. Jednakże w przypadku zasobów otwartych, nie podlegających kontroli, jak np. dokumenty tekstowe w sieci Internet, nie ma możliwości wprowadzenia tradycyjnych metod opracowania treści dokumentu przeprowadzanego przez człowieka. Dlatego systemy wyszukiwawcze repozytoriów rozległych (m.in. wyszukiwarki internetowe) dostęp do informacji organizują na zasadzie wyszukiwania pełnotekstowego. W praktyce wyszukiwanie takie nie jest przeprowadzane indywidualnie dla każdego użytkownika, a tylko jeden raz dla każdego dokumentu w momencie jego akwizycji do systemu. W tym modelu do indeksów włączane są wybrane słowa z treści. Samo zaś przygotowanie odpowiedzi na zapytanie użytkownika odbywa się na zasadzie porównania kwerendy z charakterystykami wyszukiwawczymi dokumentów.

W zakresie wyszukiwania informacji w systemach zautomatyzowanych można wyróżnić dwa główne podejścia, pomiędzy którymi różnice pojawiają się już na etapie indeksowania dokumentów tekstowych. Są to odpowiednio: reprezentowanie treści w systemach formalnych (w tym klasyfikowanie, przedmiotowanie, indeksowanie z wykorzystaniem tezaurusów) oraz indeksowanie swobodne w systemach pełnotekstowych. Starsze jest pierwsze z wymienionych podejść, czerpiące z tradycji bibliotekarskich postulujących indywidualną analizę i opis treści udostępnianego dokumentu. Ujęciu temu poświęcono kilka kolejnych akapitów.

Stosowane w systemach bibliotecznych języki informacyjno-wyszukiwawcze o kontrolowanym słownictwie, jak np. jhp KABA, zapewniają dokładny nadzór nad przypisaniem dokumentu do poszczególnych klas lub kategorii tematycznych (utworzonych przez hasła przedmiotowe proste i rozwinięte) i pozwalają zawęzić przekazywany użytkownikowi zbiór wyników. Są one stosowane z powodzeniem w zautomatyzowanych bibliotecznych systemach informacyjno-wyszukiwawczych. W początkowym okresie funkcjonowania sieci Internet najbardziej rozpowszechnionym modelem indeksowania zasobów było ich klasyfikowanie na podstawie analizy treści. Sprzyjała temu niewielka liczba źródeł oraz dokumentów dostępnych w sieci. Również pierwsze systemy informacyjne WWW funkcjonowały jako katalogi stron web. Mieczysław Kłopotek przedstawia katalogi webowe, jako jedną z najstarszych metod wyszukiwania informacji w sieci. Prezentuje je jako struktury drzewiaste, przypisujące zawartość poszczególnych witryn do określonej kategorii, przy czym poszczególne podkategorie prezentują coraz wyższy stopień szczegółowości w stosunku do kategorii nadrzędnych. Zawartość oraz struktura katalogów tworzona jest przez ludzi, dzięki czemu jest bardziej zrozumiała dla użytkownika, a odpowiedzi systemu na kwerendy są bardziej relewantne. Do wad katalogów zalicza się niski stopień aktualności w stosunku do bieżącej zawartości serwisów internetowych oraz niewielki zakres indeksowania tych zasobów[6].

Katalogi webowe stanowią klasę systemów informacyjno-wyszukiwawczych zapewniających podział zindeksowanych dokumentów na przyjęte (najczęściej *a priori*) kategorie tematyczne. W momencie wprowadzania informacji o zasobie do systemu przypisuje się go, na podstawie analizy treści, do odpowiedniej kategorii. Niewątpliwie zaletą katalogów stron WWW jest dosyć wysoki stopień relewancji dokumentów w poszczególnych kategoriach. Nie do przecenienia jest również potencjalna możliwość dokonania wstępnej oceny wartości danego dokumentu przez osobę zatwierdzającą wpis do katalogu. Kolejną korzyścią takiego sposobu indeksowania zasobów sieciowych jest wykluczenie stron z fałszywie podanym opisem zawartości[7], co przyczynia się do wzrostu wiarygodności katalogu jako źródła informacji, a pośrednio chroni użytkowników (np. przed witrynami wyłudzającymi dane osobowe).

---

[6] Por. M. Kłopotek: *Inteligentne wyszukiwarki internetowe*. Warszawa: Akademicka Oficyna Wydawnicza EXIT 2001, s. 17.

[7] Np. w polach meta dokumentu HTML: *description* i *keywords*.

Skuteczne korzystanie z takich systemów kategoryzujących wymaga jednak pewnego przeszkolenia oraz wysiłku, co siłą rzeczy ogranicza liczbę ich twórców i użytkowników. Ponadto, ze względu na kontekstowe opracowanie treści dokumentu, wymagające jej zrozumienia oraz odpowiedniej perspektywy kontekstowej i pozatreściowej, klasyfikacje czy kategoryzacje sformalizowane są trudne do automatycznego zaimplementowania w systemach komputerowych, a tylko pełna automatyzacja procesu indeksowania może zapewnić w miarę aktualną informację o dostępnych zasobach sieciowych. Ma to oczywisty wpływ na aktualność oraz kompletność informacji dostępnych w danym katalogu, które to miary są zdecydowanie niższe niż w przypadku wyszukiwarek internetowych.

Wyszukiwarki internetowe posiadają ogromną przewagę nad klasycznymi katalogami stron WWW, wynikającą z wielokrotnie większego zakresu ilościowego indeksowanych dokumentów oraz znacznie krótszego czasu opracowania i udostępnienia dokumentów. W obliczu lawinowo przyrastających zasobów Internetu możliwości największych nawet zespołów redakcyjnych katalogów są zbyt ograniczone, żeby na bieżąco zindeksować wszystkie dostępne dokumenty. Imperatyw szybkiego dostępu do najświeższej informacji wymusza zastosowanie automatyzacji w procesie wyszukiwania dokumentów w sieci oraz indeksowania ich treści. Po latach prymatu sformalizowanych systemów opracowania (będących wynikiem opracowania rzeczowego dokumentów) obserwujemy w ostatnich czasach wzrost popularności indeksowania i wyszukiwania swobodnego, za pomocą słów kluczowych. W modelu tym treść dokumentu jest reprezentowana za pomocą zbioru słów znaczących z całego tekstu. Kosztem wyszukiwania informacji za pomocą słownictwa niekontrolowanego jest duża liczba dokumentów przesłanych użytkownikowi w odpowiedzi na jego zapytanie, przy czym wiele z tych dokumentów cechuje się niską relewancją. Jednakże wydaje się, że jest to sytuacja akceptowana przez użytkowników. Na potwierdzenie tej tezy można przytoczyć słowa Jadwigi Woźniak, która stwierdza, że wysoki poziom sztuczności języków informacyjno-wyszukwiawczych powoduje, iż użytkownicy częściej korzystają z możliwości wyszukiwania poprzez słowa kluczowe. Konkluduje jednak, że podejście takie często skutkuje uzyskaniem w wyniku nadmiaru informacji[8].

[8] J. Woźniak: *Tendencje w teorii i praktyce języków informacyjno-wyszukiwawczych*. W: *Opracowanie przedmiotowe – osiągnięcia naukowe i praktyka* [on-line]. Warszawa: Wyższa Szkoła Ekonomiczno-Informatyczna 2004, s. 20. [Dostęp: 15 listopada 2011]. Dostępny w World Wide Web: http://www.wsei.pl/biblioteka/materialy/jezykinfo_ex.pdf.

Pośrednim potwierdzeniem akceptacji przez użytkowników nadmiarowości i mniejszej dokładności wyników wyszukiwania w odpowiedzi na kwerendy zbudowane ze słów kluczowych, jest sukces wyszukiwarek internetowych, które zastąpiły i w dużym stopniu wyparły systemy typu Gopher, Archie czy katalogi webowe. Dodatkowo można przywołać tu tzw. folksonomie (ang. *folksonomy*), czyli mechanizmy społecznego klasyfikowania materiałów webowych[9]. Użytkownicy serwisów Web 2.0 mogą poszczególnym dokumentom, w szerokim zakresie znaczenia tego terminu, przypisywać określenia treści za pomocą znaczników, nazywanych potocznie tagami. Znaczniki wykorzystywane są do swobodnego klasyfikowania treści dowolnych materiałów multimedialnych, nie tylko tekstu, bez jakiegokolwiek słownika regulującego dostępne hasła[10]. Oznaczanie materiałów za pomocą własnych znaczników zapoczątkował serwis delicjo.us[11]. Możliwość swobodnego opisywania treści dokumentu przez użytkowników oferuje obecnie wiele serwisów społecznościowych, np. serwis YouTube.com (http://www.youtube.com/) czy polski gwar.pl (http://www.gwar.pl)[12].

Uwzględniając trudność w tworzeniu poprawnych formalnie zapytań wyszukiwawczych (w systemach informacyjno-wyszukiwawczych), stosunkową łatwość i swobodę w przypisywaniu słów kluczowych lub ta-

---

[9] *Encyclopedia of library and information sciences* definiuje folksonomie jako języki indeksowania powstające w efekcie rozproszonegoopisywania zasobów przez liczne osoby tagujące. Por. J. Fourner: *Folksonomies*. W: *Encyclopedia of library and information sciences*, 3rd ed. Boca Raton (FL) 2010, s. 1858. Cyt. za: J. Woźniak-Kasperek: *Wiedza i język informacyjny w paradygmacie sieciowym*. Warszawa: Wydawnictwo SBP 2011, s. 193. O folksonomiach w aspekcie wyszukiwania w katalogach bibliotek por. J. Woźniak-Kasperek: *Wiedza i język...*, s. 191-196.

[10] O folksonomiach i słowach kluczowych por. m.in. P. Malak: *Słowa kluczowe i tagi jako metody swobodnego oznaczania treści dokumentów w środowisku Nowych Mediów*. W: *Zeszyty Wydziału Humanistycznego VI. Prace Medioznawcze*, pod red. A. Pawłowskiego. Jelenia Góra: Karkonoska Państwowa Szkoła Wyższa w Jeleniej Górze 2011, s. 173-187.

[11] Serwis delicio.us, początkowo: del.icio.us, (dostępny pod adresem: http://www.delicious.com/) oferuje możliwość tworzenia publicznie dostępnych zakładek do zasobów internetowych. Zasoby są kategoryzowane w sposób niehierarchiczny za pomocą znaczników (tagów).

[12] Rozwiązanie takie, jako metoda indeksowania treści dokumentów webowych, jest zdecydowanie wolniejsze niż proces automatyczny – od opublikowania danego materiału do momentu opisania jego zawartości za pomocą znaczników/tagów może upłynąć kilka dni. Jednakże jego niewątpliwą zaletą jest utworzenie opisu (będącego sumą wyrażeń zastosowanych w znacznikach) przez człowieka, z uwzględnieniem informacji pozatekstowych. Ponadto tagi są przypisywane do dowolnego typu materiałów, nie tylko do dokumentów tekstowych – za ich pomocą opisywane są również materiały multimedialne, które nie poddają się łatwo indeksowaniu automatycznemu (zdjęcia, filmy).

gów, a także rosnącą popularność tej metody oznaczania treści poszukiwanych dokumentów wśród użytkowników, można pokusić się o stwierdzenie, że techniki wyszukiwawcze bazujące na słowach kluczowych będą w najbliższym czasie dominującą formą wyszukiwania. Dlatego też wszelkie badania przybliżające praktyczne wdrożenie takich możliwości mogą przynieść bardzo cenne wyniki.

Pośrednie potwierdzenie takiego założenia znaleźć można w artykule Krzysztofa Goneta *Dlaczego słowa kluczowe a nie hasła przedmiotowe? Co dalej z opracowaniem rzeczowym w bibliotekach FIDES?* Przywoływany autor wyjaśnia przyczyny, dla których w Katalogu Centralnym Bibliotek FIDES oraz w multiwyszukiwarce FIDKAR zastosowano do charakterystyki treści dokumentów właśnie słowa kluczowe, rezygnując z wykorzystania jhp. Według niego na taki stan rzeczy wpłynęły przyczyny formalne oraz praktyczne. Przeszkodą formalną zastosowania jhp w katalogach FIDES był brak odpowiedniego wzorca haseł przy jednocześnie zbyt wysokim poziomie ogólności haseł już istniejących. Ponadto w bibliotekach kościelnych problemem okazał się brak bibliotekarzy przeszkolonych w stosowaniu języka haseł przedmiotowych. Kolejną, istotną przyczyną jest uogólniający charakter tego typu języka informacyjno-wyszukiwawczego w porównaniu do słów kluczowych czy tezaurusów. W zakresie praktycznych rozwiązań zaś, słowa kluczowe, a jeszcze lepiej deskryptory, umożliwiają tworzenie złożonych zapytań wyszukiwawczych. Każdy dokument w zbiorach FIDES opisywany jest za pomocą wielu wyrażeń kluczowych reprezentujących jego treść. Połączenie kilku słów kluczowych w zapytaniu użytkownika pozwala w wyraźny sposób zawęzić zbiór wyników wygenerowany przez system, przy zwiększeniu stopnia zgodności wyszukanych dokumentów z zadanym zapytaniem[13].

Przy okazji rozważań nad indeksowaniem formalnym oraz swobodnym warto wspomnieć o interesujących badaniach przeprowadzonych przez Ewę Głowacką. Badaczka ta przeprowadziła eksperyment mający ocenić dokładność i kompletność wyszukiwania. Badania prowadzone były metodą F. W. Lancastera dla zapytań informacyjnych wyrażonych w dwóch wersjach języka haseł przedmiotowych (jhp Biblioteki Narodowej: hasła przed-

[13] K. Gonet: *Dlaczego słowa kluczowe a nie hasła przedmiotowe? Co dalej z opracowaniem rzeczowym w bibliotekach FIDES?* „FIDES – Biuletyn Bibliotek Kościelnych" [on-line] 2004, nr 1-2 (18-19), s. 24 i nast. [Dostęp: 19 kwietnia 2011]. Dostępny w World Wide Web: http://digital.fides.org.pl/dlibra/docmetadata?id=29&from=&dirids=1&ver_id=4698&lp=1&QI=!94F2E9847571EDDCB7D673E25382F1B2-7.

miotowe w formie tradycyjnej – temat i określniki traktowane jako jedno wyrażenie oraz w formie elastycznej – temat i określniki osobno), a także w języku słów kluczowych. Zwiększenie elastyczności hasła przedmiotowego przyczyniło się do wzrostu jego efektywności, przy czym jhp oferował wyższą kompletność, zaś język słów kluczowych wyższą dokładność[14].

Charakterystyki wyszukiwawcze mogą być przygotowane zgodnie z różnymi wytycznymi, jednakże wśród użytkowników zasobów internetowych najpopularniejszą metodą reprezentowania w zapytaniach treści dokumentów są wyrażenia kluczowe, generowane na podstawie słownictwa niekontrolowanego. Słowa kluczowe odwzorowujące treść dokumentu tworzą jego charakterystykę słowną. W tradycyjnej analizie dokumentów charakterystyki słowne powstają w trakcie procesu opracowania rzeczowego. Podczas analizy treści dokumentu prowadzonej przez człowieka wskazuje się wyrażenia charakteryzujące treść danej pozycji. Tradycyjne systemy informacyjno-wyszukiwawcze operują na reprezentacjach treści dokumentów, do których dostęp potrzebny jest użytkownikom. Przygotowanie prawidłowej reprezentacji, wyrażonej w odpowiednim, stosowanym w danym systemie języku informacyjno-wyszukiwawczym, wymaga zarówno wprawy, ze względu na konieczność opanowania leksyki i gramatyki danego języka, jak i czasu, z powodu wymogu zapoznania się z treścią dokumentu i przygotowaniem odpowiedniego opisu owej treści. Możliwość zautomatyzowania tego procesu niesie ze sobą niewątpliwe korzyści w postaci oszczędności kosztów pracy oraz czasu związanych z opracowaniem odpowiedniej charakterystyki, co mają m.in. wskazać badania podjęte w niniejszej książce.

Cel i struktura pracy

Za cel badawczy niniejszej książki postawiono porównanie skuteczności metod automatycznych i kognitywnych w tworzeniu charakterystyk wyszukiwawczych dokumentów za pomocą słów kluczowych. Ponadto założono przeprowadzenie badania i oceny możliwości automatycznego generowania słów kluczowych jako reprezentacji treści dokumentów. W badaniach i w powstałej na ich podstawie książce posłużono się metodą analizy i krytyki piśmiennictwa oraz metodami statystycznymi. Badania własne zostały

[14] Por. E. Głowacka: *Badania efektywności języków informacyjno-wyszukiwawczych (komunikat z badań)*. W: *Komputeryzacja bibliotek. Materiały konferencji 24-26 maja 1993 r.*, pod red. B. Ryszewskiego. Toruń: Wydawnictwo UMK 1994, s. 209-210.

przeprowadzone z wykorzystaniem autorskiego systemu analizy kwantytatywnej tekstów języka polskiego, przy użyciu metod statystycznych do ustalenia i analizy frekwencji wyrażeń językowych w korpusie tekstów.

Całość została podzielona na dwie części. Część pierwsza, składająca się z trzech rozdziałów (*Wstęp* i dwa rozdziały), poświęcona jest teorii badań nad przetwarzaniem języka naturalnego, z zawężeniem do tematyki niniejszej książki. Pierwszy rozdział, zatytułowany *Związki NLP z informacją naukową*, prezentuje podstawy teoretyczne omawianej dziedziny badawczej[15]. Zostały w nim zaprezentowane cele oraz geneza tego rodzaju działalności badawczej. Przedyskutowano również problemy nazewnicze związane z tą dziedziną.

W rozdziale drugim, *Ustalenia terminologiczne oraz wybrane metody komputerowego przetwarzania języka naturalnego*, przedstawiono stosowaną w książce nomenklaturę oraz wybrane strategie nauki o przetwarzaniu tekstów języka naturalnego. Zaprezentowane zostały także jednostki badania kwantytatywnego tekstów oraz cechy statystyczne jednostek leksykalnych. Omówiono również wybrane, przydatne w badaniach kwantytatywnych, sposoby reprezentacji treści dokumentów. Oprócz stosowanego na potrzeby niniejszej książki pojęcia **wielozbioru** opisane zostały także **lista frekwencyjna** i **model wektorowy**. W rozdziale tym zostały również zaprezentowane metody optymalizacji tekstu na potrzeby automatycznego przetwarzania.

Druga część książki, składająca się z trzech rozdziałów (dwa rozdziały oraz *Podsumowanie*) opisuje przebieg i wyniki badań własnych przeprowadzonych na potrzeby niniejszej publikacji. W rozdziale trzecim, *Zasady postępowania badawczego i opis przygotowanego systemu*, przedstawione zostały zasady, według których przeprowadzono badania oraz system stworzony na potrzeby niniejszej książki. Zaprezentowany został szczegółowo cel oraz przedmiot badań, a także założone hipotezy badawcze. W rozdziale tym omówiłem również organizację i przebieg badań, zamieściłem także opis materiału badawczego.

Wyniki analiz oraz porównanie analizowanych w pracy metod wyznaczania słów kluczowych zostały zaprezentowane w rozdziale czwartym,

[15] Pod pojęciem przetwarzanie języka naturalnego rozumiana jest automatyczna realizacja tego procesu. W dalszej części książki oba wyrażenia: *przetwarzanie języka naturalnego* oraz *automatyczne (komputerowe) przetwarzanie języka naturalnego* będą traktowane wymiennie, jako równoznaczne.

Analiza oraz interpretacja materiału badawczego i wyników badań. Dokonałem w nim analizy zastosowanych metod ustalania wagi leksemów oraz wpływu tych metod na generowane automatycznie listy słów kluczowych.

Podsumowanie przedstawia wnioski sformułowane na podstawie wyników badań. Zostały w nim omówione problemy, które pojawiały się w trakcie badań oraz przygotowań. Zaproponowałem również możliwości praktycznego wykorzystania wyników uzyskanych w rezultacie prac badawczych, a także kierunki dalszego rozwoju badań w zakresie automatycznego przygotowywania charakterystyk treści dokumentów.

Pozostaje wierzyć, że tak skonstruowana całość okaże się przydatna zarówno jako pomoc podczas zajęć poświęconych wyszukiwaniu informacji, jak i dla osób pracujących z informacją cyfrową, bibliotekarzy oraz badaczy procesów przetwarzania informacji. Usprawnienie i podniesienie jakości procesu wskazywania relewantnych słów kluczowych wychodzi naprzeciw oczekiwaniom i przyzwyczajeniom użytkowników informacji cyfrowej. Książka, z założenia obejmująca zaawansowane zagadnienia wyszukiwania informacji, jest kierowana do badaczy i pracowników sektora informacyjnego. Szczególnie przydatna może być dla osób współtworzących biblioteki cyfrowe. Wyniki badań dotyczących możliwości automatycznego generowania słów kluczowych charakteryzujących treść dokumentu mogą okazać się przydatne we wszelkiego rodzaju repozytoriach cyfrowych.

W niniejszej książce wykorzystałem wyniki badań przeprowadzonych na potrzeby rozprawy doktorskiej. Zarówno rozprawa, jak i książka nie powstałyby, gdyby nie cenna pomoc wielu osób, z których kilku chciałbym podziękować w wyjątkowy sposób. Szczególne podziękowania chciałbym złożyć promotorowi rozprawy, profesorowi Adamowi Pawłowskiemu, który nieustannie wspomagał mnie podczas pracy i pomagał w dotrzymywaniu terminów.

Dziękuję serdecznie pracownikom Instytutu Informacji Naukowej i Bibliologii UMK w Toruniu, w szczególności profesor Bronisławie Woźniczce-Paruzel za udzielone wsparcie i wiarę w sukces oraz profesorowi Januszowi Tondelowi.

Książka w obecnej postaci zawdzięcza wiele szanownym recenzentom, profesor Irenie Kamińskiej-Szmaj, profesor Jadwidze Woźniak-Kasperek oraz profesorowi Wiesławowi Babikowi.

Chciałbym również podziękować rodzinie, za cierpliwość i wsparcie podczas pracy.

Związki NLP z informacją naukową

Analiza kwantytatywna jest elementem lingwistyki i początkowo była stosowana wyłącznie do tekstów w celu wykrycia prawidłowości językowych. Obecnie wykorzystywana jest również w informacji naukowej, chociażby przy indeksowaniu automatycznym. Współczesna nauka o informacji, w odniesieniu do rozległych kolekcji dokumentów, odwołuje się, w mniejszym bądź większym stopniu, do osiągnięć i zdobyczy przetwarzania języka naturalnego. Warto przy okazji omawiania możliwości wykorzystania rozwiązań inżynierii lingwistycznej wprowadzić niezbędną nomenklaturę NLP[16].

1.1. Wprowadzenie teoretyczne do przetwarzania języka naturalnego

Pomimo stosunkowo bogatej historii wykorzystania komputerów do przetwarzania tekstów języka naturalnego ta dziedzina aktywności nie doczekała się dotychczas zwięzłej, precyzyjnej i jednoznacznej definicji. Najczęściej przedstawiana jest w sposób opisowy, poprzez zaprezentowanie ogólnego celu, omówienie prowadzonych badań, bądź też wska-

[16] O historii badań nad przetwarzaniem języka naturalnego por. m.in.: A. Mykowiecka: *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym*. Warszawa: Wydawnictwo PJWSTK 2007. Fragmenty niniejszego rozdziału można znaleźć w: P. Malak: *Rozwój badań nad przetwarzaniem języka naturalnego*. „Zagadnienia Informacji Naukowej" 2010, nr 2 (96), s. 21-30.

zanie innych dyscyplin, na których pograniczu można ją usytuować. Nie ma również jednoznacznie przyjętej nazwy, w zależności od kontekstu badań poszczególni badacze, w tym również polscy, jak np. Maciej Piasecki, Adam Przepiórkowski czy Agnieszka Mykowiecka, stosują różne określenia dla prac związanych z automatycznym przetwarzaniem danych językowych[17]. Niniejszy rozdział ma na celu wprowadzenie do teorii dyscypliny oraz uporządkowanie terminologii.

1.2.1. Ustalenia terminologiczne związane z nazwą badań nad tekstami języka naturalnego

Otwartym pozostaje problem nazwy dla omawianej dziedziny. Zarówno w piśmiennictwie zagranicznym, jak i polskim można spotkać kilka terminów traktowanych jako równoznaczne, stosowanych w zależności od głównego celu prowadzonych prac. Wspominani wcześniej badacze proponują, za literaturą anglojęzyczną, następujące określenia, zaznaczając równoznaczność użytych terminów:

- przetwarzanie języka naturalnego (ang. *natural language processing*, NLP) – termin najstarszy, najpowszechniej stosowany,
- inżynieria lingwistyczna (ang. *language engineering*, LE) – przy czym wydaje się, że termin ten jest obecnie najpopularniejszy wśród polskich badaczy,
- lingwistyka komputerowa lub lingwistyka informatyczna (ang. *computational linguistic*, CL),
- inżynieria języka naturalnego (ang. *natural language engineering*, NLE),
- technologia języka (ang. *language technology*, LT lub *human language technology*, HLT) – termin raczej wychodzący z użycia, wskazuje na silne związki prowadzonych prac z informatyką stosowaną, co nie jest zgodne ze stanem faktycznym[18].

Interesującą analizę nazwy dla terminu **lingwistyka informatyczna** przeprowadził M. Piasecki w publikacji przytaczanej już we wcześniej-

[17] O problemach związanych z jednoznacznym zdefiniowaniem dziedziny oraz terminologii por. m.in.: A. Mykowiecka: *Inżynieria lingwistyczna...*, s. 9 i 14; A. Przepiórkowski: *Powierzchniowe przetwarzanie języka polskiego*. Warszawa: Akademicka Oficyna Wydawnicza EXIT 2008, s. 3-6; M. Piasecki: *Cele i zadania lingwistyki informatycznej*. W: *Metodologie językoznawstwa. Współczesne tendencje i kontrowersje*, pod red. P. Stalmaszczyka. Kraków: Lexis 2008, s. 252-254. [Dostęp: 19 września 2011] Dostępny w World Wide Web: http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/MetJezII-piasecki-ostateczna.pdf.

[18] Por. m.in.: tamże, s. 14; A. Przepiórkowski: dz. cyt., s. 5.

szej części niniejszej książki. Poniżej zostanie zaprezentowane streszczenie owych rozważań oraz wynikające z nich wnioski.

Problemem budzącym wątpliwości cytowanego badacza jest kwestia przymiotnika użytego w nazwie: **informatyczna** czy **komputerowa**. Należy tu nadmienić, że oba pojęcia (lingwistyka informatyczna oraz lingwistyka komputerowa) funkcjonują równolegle w polskim piśmiennictwie przedmiotu. Wydaje się, że użycie konkretnego przymiotnika dyktowane jest chęcią wykazania maszynowego aspektu przetwarzania danych językowych (dla terminu *komputerowa*) lub też wskazania na automatyczne, aplikacyjne przetwarzanie wyrażeń języka naturalnego (termin *informatyczna*). Historycznie termin polski ewoluował od lingwistyki komputerowej do informatycznej. Powołując się na źródłosłów terminu, angielskie określenie *Computational Linguistics*, w którym pierwszy człon wskazuje na metody obliczeniowe, oraz odwołując się do polskiej tradycji używania słowa informatyka, M. Piasecki optuje za wersją **lingwistyka informatyczna**, jako bliższą istoty sprawy[19].

Ponieważ termin lingwistyka informatyczna można interpretować jako modelowanie języka za pomocą komputerów, traktowanych w tym przypadku wyłącznie jako zaawansowane narzędzia obliczeniowe[20], w literaturze polskiej (za lit. anglojęzyczną) dla określenia prac związanych z komputerowym przetwarzaniem języka naturalnego używa się często pojęcia **inżynieria lingwistyczna**. Termin ten oznacza wykorzystanie wyników badań NLP do tworzenia aplikacji wykorzystujących dane językowe. Pojawia się on m.in. w pracach cytowanych wcześniej badaczy: A. Mykowieckiej czy A. Przepiórkowskiego.

Termin **przetwarzanie języka naturalnego**, będący historycznym określeniem tej dziedziny aktywności, jest najbardziej uniwersalnym, ale też i najbardziej ogólnym. Wraz z rozwojem dziedziny pojawiały się kolejne określenia, zawężające pojęcie, doprecyzowujące cele bądź umiejscowienie w stosunku do innych dziedzin. Jednakże nadal często prace związane z komputerowym przetwarzaniem tekstów języka naturalnego określa się przetwarzaniem języka naturalnego lub skrótem NLP. Pozostałe terminy wskazują na bliższe związki badań z lingwistyką i modelowaniem języka,

[19] Por. M. Piasecki: dz. cyt., s. 253.

[20] W takim znaczeniu terminu lingwistyka informatyczna używa m.in. Marek Świdziński; por. M. Świdziński: *Lingwistyka korpusowa w Polsce – źródła, stan, perspektywy* [on-line]. „LingVaria" 2006, nr 1, s. 25. [Dostęp: 19 września 2011]. Dostępny w World Wide Web: http://www2.polonistyka.uj.edu.pl/LingVaria/archiwa/LV_1_2006_pdf/02_swidzinski.pdf.

z wykorzystaniem komputerów do organizowania, przetwarzania i analizowania danych lingwistycznych, bądź też, jak np. inżynieria lingwistyczna, z komputerową implementacją wyników badań lingwistycznych. W niniejszej pracy, ze względu na jej aspekt programistyczny, implementujący efekty badań nad opisem języka, dla określenia prac związanych z komputerowym przetwarzaniem tekstów języka naturalnego będą stosowane zamiennie oba terminy: **przetwarzanie języka naturalnego** (NLP) oraz **inżynieria lingwistyczna** (LE).

1.2. Wybrane kierunki badawcze

Poniżej zostaną zaprezentowane wybrane i sprawdzone metody NLP związane z tematyką niniejszej książki. Należą one do nurtu statystycznego, który, przypomnijmy, cechuje się stosunkowo niskimi kosztami operacyjnymi analizy. Jednym z najstarszych zastosowań automatycznego przetwarzania danych językowych jest wyszukiwanie informacji w dokumentach tekstowych; kolejnym, wnoszącym przydatne rozwiązania, jest grupowanie dokumentów.

1.2.1. Wyszukiwanie informacji w dokumentach

Ch. Manning i pozostali autorzy wyróżniają, jako element szerszego nurtu IR (ang. *Information Retrieval*), pełnotekstowe wyszukiwanie informacji z wykorzystaniem słownictwa niekontrolowanego i przeciwstawiają je modelowi wyszukiwania strukturalnego, stosowanemu najczęściej w bazach danych (m.in. relacyjnych), bądź np. w zautomatyzowanych katalogach bibliotecznych. Wyszukiwanie w zbiorach informacji strukturalnej wymaga znajomości struktury wykorzystanej do przechowywania danych, przeznaczenia poszczególnych pól oraz powiązań (relacji) zachodzących pomiędzy różnymi elementami rekordu. Przede wszystkim jednak, konieczne jest pracochłonne przygotowanie takiej struktury, podczas gdy zwykłe dokumenty tekstowe są powszechnie dostępne. Proces wyszukiwania polega m.in. na wskazaniu pola, którego zawartość ma zostać porównana do zapytania oraz wyborze sposobu bądź metody porównawczej. Konieczność posiadania wiedzy o działaniu systemu implikuje wymóg przeszkolenia użytkownika, co ogranicza możliwości skutecznego wyszukiwania w takim systemie. Natomiast metodologia wyszukiwania pełnotekstowego zakłada w pełni swobodne przeszukiwanie całego tek-

stu oraz ewentualnych metainformacji dotyczących analizowanego dokumentu. Zapytania w tym przypadku budowane są zazwyczaj w postaci listy słów bądź wyrażeń kluczowych opisujących informacje, na których zależy użytkownikowi. Wynikiem takiego wyszukiwania jest lista dokumentów zawierających jednostki leksykalne wskazane w zapytaniu. W przypadku wyszukiwania pełnotekstowego zbiór dokumentów do przeszukania nie musi zostać wstępnie opracowany przez człowieka – przeglądanie i porównywanie zawartości dokumentów tekstowych jest procesem w pełni zautomatyzowanym. Takie podejście wyróżnia się dwiema ważnymi cechami. Po pierwsze, jest łatwe i wygodne dla użytkownika, który w całkowicie dowolny sposób może wskazywać interesujące go słowa kluczowe z treści tekstu. Po drugie zaś, wyszukiwanie pełnotekstowe pozwala na obniżenie kosztów samego procesu przetwarzania dokumentów poprzez pominięcie m.in. etapu opracowania rzeczowego, a czasem także formalnego danego dokumentu przez człowieka. Dzięki temu nowy dokument jest dostępny dla użytkowników od razu po wprowadzeniu jego treści do systemu, bez opóźnienia wynikającego z konieczności opracowania rzeczowego.

Na korzyść słów kluczowych jako formy komunikacji użytkownika z systemem można powołać się, za Barbarą Sosińską-Kalatą, na wyniki badań nad kognitywnym podejściem do projektowania systemów informacyjnych. Autorka przytacza w pracy *Modele organizacji wiedzy w systemach wyszukiwania informacji o dokumentach* wyniki badań nad zależnościami między wiedzą użytkownika o systemie (dalej WS) a rezultatami wyszukiwania informacji[21]. Z podanych informacji można wywnioskować, że we wczesnym okresie funkcjonowania takich systemów (lata 80.) WS była istotnym elementem wpływającym na jakość wyszukiwania. Natomiast w miarę ewoluowania systemów i ich interfejsów WS odgrywała coraz mniej znaczącą rolę. Szczególną uwagę warto zwrócić na opis eksperymentu podany w przypisie 6 na stronie 27 cytowanej publikacji. Opisuje on eksperyment przeprowadzony na dwóch grupach studentów wyszukujących informacje w pełnotekstowych bazach danych na CD ROM. Obie grupy zostały przeszkolone w różnym stopniu w zakresie strategii wyszukiwania, natomiast wyniki ich wyszukiwań były podobne. Środowisko pełnotekstowe pozwalało użytkownikom na wyszukiwanie za pomocą

[21] Za: B. Sosińska-Kalata: *Modele organizacji wiedzy w systemach wyszukiwania informacji o dokumentach*. Warszawa: SBP 1999, s. 27.

słownictwa niekontrolowanego, czyli wskazywania dowolnych słów kluczowych występujących w treści dokumentu[22].

Ze spostrzeżeniami tymi koresponduje definicja współczesnego użytkownika informacji, jaką podaje w swojej książce Mieczysław A. Kłopotek. Na poparcie swej oceny podaje następujące cechy typowego użytkownika:

- brak gotowości do zapoznania się z instrukcjami obsługi (tu: wyszukiwania),
- niedokładne zapytania, zwracające bardzo liczne zbiory dokumentów,
- przeglądanie najczęściej kilku pierwszych wyników z listy, w związku z czym niezbędne jest odpowiednie ustalanie relewancji poszczególnych dokumentów,
- oczekiwanie pomocy w znalezieniu odpowiedzi,
- oczekiwanie szybkiego uzyskania adekwatnych wyników[23].

Założenia pełnotekstowego wyszukiwania informacji z wykorzystaniem słownictwa niekontrolowanego zostały w praktyce zaimplementowane m.in. w wyszukiwarkach internetowych. Intuicyjne interfejsy użytkownika wyszukiwarek oraz możliwość wskazania dokumentów jedynie na podstawie słów kluczowych występujących w treści pozwoliły na swobodne prowadzenie wyszukiwania informacji przez miliony użytkowników Internetu[24].

Kolejna oszczędność kosztów operacyjnych, a jednocześnie racjonalizacja procesu wyszukiwania pełnotekstowego, została osiągnięta po wprowadzeniu plików indeksów odwróconych. W plikach tych przechowywana jest informacja o lokalizacji każdego wystąpienia tokenu bądź słowa we wszystkich dokumentach kolekcji. Adresy przypisane do poszczególnych kluczy w plikach indeksowych mogą, w zależności od systemu, zawierać identyfikatory plików, w których przechowywana jest treść dokumentów lub identyfikatory plików wraz z informacjami lokalizującymi dany klucz w konkretnym pliku. Proces odwzorowania lokalizacji poszczególnych słów lub tokenów nazywany jest procesem indeksowania. Przeprowadzany jest on w momencie akwizycji dokumentu do zbioru. W związku z takim zorganizowaniem informacji, system wyszukiwawczy może odwołać się bezpośrednio do wskazanego przez użytkownika słowa lub wyrażenia kluczowego i w krótkim czasie wskazać wszystkie dokumenty zawierające dane wyrażenie bez konieczności każdorazowego analizowania treści dokumen-

[22] Por. tamże, s. 27.
[23] Por. M. Kłopotek: dz.cyt, s. 24.
[24] Ch. D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 2-3.

tów dla poszczególnych zapytań od użytkowników. Dla ułatwienia działania klucze przechowywane w plikach indeksów odwróconych posortowane są według kolejności alfabetycznej. Wyrażenia z zapytania użytkownika porównywane są z zawartością plików indeksów odwróconych. Alfabetyczny układ kluczy w tych plikach skraca czas dotarcia do szukanej informacji, nie ma potrzeby porównywania zapytań z treścią dokumentów, system wyszukiwawczy odwołuje się od razu do odpowiedniego fragmentu indeksu odwróconego i pobiera dane lokalizujące konkretne teksty.

Wyszukiwanie takie nie jest jednak wolne od wad. J. Woźniak-Kasperek w książce *Wiedza i język informacyjny w paradygmacie sieciowym* podaje następujące, wybrane mankamenty wyszukiwania pełnotekstowego:

- brak kontroli synonimii,
- brak rozwiązania problemu nazw pełnych i ich skrótów,
- brak możliwości równoległego wyszukiwania wyrażeń podanych w różnych językach,
- brak powiązań pomiędzy wyrażeniami współczesnymi a przestarzałymi,
- problem wyrażeń homonimicznych,
- różnice w algorytmach tworzenia stop list[25].

Za cytowaną już pracą M. Kłopotka można wymienić następujące rodzaje wyszukiwania:

- według słów kluczowych (wskazywanie dokumentów zawierających jedno lub więcej ze słów podanych przez użytkownika),
- boolowskie (z zastosowaniem operatorów logicznych),
- koncepcyjne (wskazanie dokumentów związanych znaczeniowo z podanymi przez użytkownika słowami),
- szukanie frazy (wskazanie dokumentów zawierających podany przez użytkownika ciąg wyrazów),
- szukanie z określeniem odległości słów (z podaniem odległości, w jakiej powinny znajdować się w dokumentach słowa podane przez użytkownika),
- za pomocą tezaurusa (automatycznie wykorzystywany przez wyszukiwarkę zbiór synonimów),
- wyszukiwanie rozmyte (ang. *fuzzy search*, wyszukiwanie dopasowania również części słów, np. samych tematów bez końcówek),
- szukanie dokumentów podobnych[26].

[25] Por. J. Woźniak-Kasperek: dz.cyt., s. 188-191.
[26] Za: M. Kłopotek: dz. cyt., s. 178.

Spośród wskazanych rodzajów można wyróżnić dwa podstawowe podejścia do wyszukiwania informacji: **model oparty na logice Boole'a** (ang. *Boolean Logic Model*, BLM) oraz **model rankingowy** (ang. *ranked-output model*). W przypadku modelu logiki dwuwartościowej zapytanie buduje się ze słów bądź fraz połączonych operatorami logicznymi. Metoda ta pozwala wyłonić ze zbioru dokumentów te, których treść spełnia zadany warunek. Jest więc podejściem zero-jedynkowym (dokumenty należą lub nie do wyselekcjonowanego zbioru). W modelu tym dokument jest reprezentowany jako zbiór słów, zaś kwerendy prezentowane są jako wyrażenia boolowskie z wykorzystaniem spójników logicznych (AND, OR, NOT) oraz nawiasów, za pomocą których określa się kolejność operacji. Ze względu na łatwość implementacji oraz kontrolę nad wynikami (osiąganą za pomocą złożonych wyrażeń logicznych), model ten jest najpopularniejszym spośród modeli wyszukiwania.

Wadą modelu boolowskiego jest jego mała elastyczność, ze względu na specyfikę operatorów logicznych można wskazać jedynie dokumenty zawierające każde z zadanych (operator AND) haseł lub tylko jedno z nich (operator OR). Oczywiście operator alternatywy pozwala wskazać dokumenty zawierające kilka z zadanych haseł, ale nie daje możliwości wskazania ile z tych haseł powinno wystąpić w dokumencie, aby wynik został uznany za trafny. W modelu tym nie można stosować rankingów, ponieważ wszystkie dokumenty pasujące do kwerendy spełniają ją w równym stopniu. Nie istnieje również możliwość łatwego doprecyzowania kwerendy (za pomocą operatora AND czy OR)[27].

Alternatywny model rankingowy pozwala ocenić podobieństwo treści dokumentów z treścią zapytania i utworzyć na tej podstawie listę rankingową dokumentów trafnych. Przy tworzeniu rankingów wykorzystywane są najczęściej następujące modele oceny podobieństwa[28]:

1. Model wektorowy (ang. *Vector Space Model*, VSM)[29].

---

[27] Tamże, s. 179.

[28] Por. A. Kempa: *Zastosowanie rozszerzonej metodologii wnioskowania na podstawie przypadków – Textual CBR w pracy z dokumentami tekstowymi* [on-line]. Systemy wspomagania decyzji. Archiwum publikacji. Katowice: Akademia Ekonomiczna, Katedra Informatyki 2003-2010. [Dostęp: 19 września 2011]. Dostępny w World Wide Web: http://www.swo.ae.katowice.pl/_pdf/221.pdf.

[29] O VSM por. G. Salton, A. Wong, C. S. Yang: *A vector space model for automatic indexing* [on-line]. „Communications of the ACM" 1975, vol. 18, nr 11, s. 613-620. [Dostęp: 25 września 2011]. Dostępny w World Wide Web: http://openlib.org/home/krichel/courses/lis618/readings/salton75.pdf. W modelu tym zarówno dokument, jak i kwerenda opisane są za pomocą terminów indeksujących. M. Kłopotek podaje, że indeksowanie fraz jest reprezentacją tekstu o funkcjonalności niższej niż indeksowanie pojedynczych słów (reprezentacja prosta). – Por. M. Kłopotek: dz. cyt., s. 179.

2. Model probabilistyczny (ang. *Probabilistic Model*, PM).

Obie metody można połączyć, wyszukując za pomocą algebry Boole'a dokumenty zgodne z zapytaniem, a następnie, oceniając stopień zgodności, przedstawić je użytkownikowi w postaci listy rankingowej.

1.2.2. Grupowanie dokumentów (klasteryzacja)

Celem grupowania dokumentów, zwanego klasteryzacją, jest wyodrębnienie na podstawie treści zbioru grup dokumnetów jednorodnych pod względem pewnej cechy treściowej. W wyniku tego procesu otrzymuje się podzbiory dokumentów podobnych do siebie, a różniących się od dokumentów w pozostałych podzbiorach. Kryteria stosowane podczas operacji kategoryzacji są kryteriami formalnymi, związanymi z występowaniem poszczególnych słów w treści dokumentów i wyznaczanym na tej podstawie podobieństwie dokumentów[30].

Autorzy pracy *An introduction to information retrieval* opisują grupowanie jako najpowszechniejszą formę uczenia się nienadzorowanego – zdobywanie wiedzy przez systemy komputerowe bez udziału człowieka. Przypisanie dokumentu do wybranej klasy odbywa się automatycznie, podobnie zresztą, jak wskazanie klas tematycznych dla analizowanego zbioru dokumentów. Kategoryzację przeciwstawiają klasyfikacji, w której dokumenty przypisywane są do klas ustalonych *a priori*. W związku z czym klasyfikacja nazywana jest uczeniem się nadzorowanym[31].

Podstawą wskazywania klas i przypisywania do nich dokumentów jest wyznaczenie odległości między dokumentami w przestrzeni dwuwymiarowej[32]. Metody wyznaczania odległości związane są z reprezentacją treści dokumentów w postaci wektorów i zostały omówione w rozdziale poświęconym technikom i metodom NLP.

---

[30] Grupowaniem wyników wyszukiwania interesuje się m.in. Dawid Weiss, por. D. Weiss, J. Stefanowski: *Web search results clustering in Polish: experimental evaluation of Carrot*. [on-line]. [Dostęp: 19 września 2011]. Dostępny w World Wide Web: http://www.cs.put.poznan.pl/dweiss/site/publications/download/iipwm-dweiss-2003.pdf; S. Osiński, J. Stefanowski, D. Weiss: *Lingo: search results clustering algorithm based on singular value decomposition* [on-line], s. 7-8. [Dostęp: 19 września 2011]. Dostępny w World Wide Web: http://www.cs.put.poznan.pl/dweiss/site/publications/download/iipwm-osinski-weiss-stefanowski-2004-lingo.pdf.

[31] Ch. D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 349.

[32] Porównanie odległości poprzedzone jest redukcją wymiarów wektorów reprezentujących treść dokumentów.

Pomimo dosyć wyraźnego rozróżnienia pól znaczeniowych klasyfikacji i kategoryzacji, nadal wielu autorów utożsamia te pojęcia. W pracy *Klasyfikacja. Narzędzia zarządzania i wyszukiwania informacji* autor zauważa, że pomimo dużej popularności grupowania i kategoryzacji pojęcia te są ze sobą mylone. Jako przykład podaje publikację *Electronic Statistics Textbook,* w której grupowanie (ang. *cluster analysis*) opisane jest jako metoda obejmująca algorytmy klasyfikacji (ang. *classification*)[33]. Sam zaś definiuje system kategoryzacji dokumentów (ang. *document categorisation* lub *document classification*) jako system przyporządkowujący dokument do jednej lub wielu uprzednio zdefiniowanych klas. Na potrzeby kategoryzacji klasy definiowane są na podstawie zbioru trenującego, czyli dokumentów zaklasyfikowanych tradycyjnie, np. przez ekspertów. Natomiast grupowanie (ang. *document clustering*) jest procesem polegającym na automatycznym wskazywaniu klas i przypisywaniu do nich dokumentów bez dostarczonych wzorców. Oba procesy są przykładami uczenia maszynowego, przy czym kategoryzacja/klasyfikacja jest przykładem uczenia z nadzorem, zaś grupowanie jest przykładem uczenia bez nadzoru[34].

Również B. Sosińska-Kalata częściowo łączy pola znaczeniowe obu terminów, przyjmując jako kategorię i kategoryzację klasyfikacje sztuczne, zdefiniowane poprzez cechę dystynktywną ważną ze względu na założone cele praktyczne klasyfikacji. Wyróżnia przy tej okazji dwa podstawowe rodzaje kategoryzacji. Jako pierwszy wskazuje porządkowanie obiektów w grupy identyfikujące ich ogólny wzorzec, w ogólne klasy (zwane przez autorkę kategoriami). Drugim rodzajem jest porządkowanie klas w superklasy, klasy ogólniejsze[35].

J. Woźniak zauważa słusznie, że pojęcie kategoryzacji jest wieloznaczne, wskazując że termin ten może oznaczać procesy wyodrębniania obiektów i ich zbiorów ocenionych jako jednorodne według danego kryterium (czyli tworzenie kategorii) lub proces przypisywania obiektów do istniejących kategorii (oznaczany również jako kategoryzowanie)[36]. Ponadto stwierdza, że również w teorii i praktyce języków informacyjno-wyszuki-

---

[33] *Electronic Statistics Textbook* [on-line]. Tulsa: StatSoft, Inc. 2001. [Dostęp: 11 stycznia 2012]. Dostępny w World Wide Web: http://www.statsoft.com/textbook/, cyt. za: P. Gawrysiak, dz. cyt., s. 13.

[34] P. Gawrysiak: dz. cyt., s. 13-14.

[35] B. Sosińska-Kalata: *Klasyfikacja. Struktury organizacji wiedzy, piśmiennictwa i zasobów informacyjnych*. Warszawa: Wydawnictwo SBP 2002, s. 22-25.

[36] J. Woźniak: *Kategoryzacja. Studium z teorii...*, s. 65.

wawczych termin kategoryzacji nie jest jednoznaczny. Pod tym pojęciem rozumiane są zarówno procesy tworzenia kategorii, procesy przyporządkowywania do istniejących kategorii obiektów systemu informacyjno-wyszukiwawczego, jak i system kategorii w jiw[37].

Cytowana badaczka zauważa, że terminy kategoryzacja i klasyfikacja często są ze sobą utożsamiane i stosowane wymiennie również w publikacjach dotyczących systemów informacyjno-wyszukiwawczych. Przy czym autorka konstatuje, że zazwyczaj wymienność użycia jest jednokierunkowa, terminy klasyfikowanie i klasyfikacja traktowane są jako synonimy terminów kategoryzowanie i kategoria. Tę trudność jednoznacznego przyjęcia zakresu znaczeniowego poszczególnych terminów tłumaczy m.in. faktem niejednoznaczności pojęcia klasyfikacja. Autorka definiuje ją jako wielostopniowy podział logiczny ograniczony dodatkowymi warunkami pozaformalnymi. Kategoryzacja umożliwia przypisanie jednego obiektu do kilku grup, zaś klasyfikacja jest przyporządkowaniem jednoznacznym i wyłącznym[38].

Kategoryzacja, jako działanie języka informacyjno-wyszukiwawczego i jego rezultat wykazuje podobieństwo semantyczne z indeksowaniem. Jako argument przeciwko synonimiczności obu terminów autorka podaje przykład języka słów kluczowych, który cechuje się niekategorialną strukturą pola semantycznego, a ewentualna kategoryzacja, w tym jiw, ma charakter niejawny[39].

Badacze zagadnienia wskazują dwa poziomy grupowania: **grupowanie płaskie** oraz **grupowanie hierarchiczne**. Pierwszy rodzaj tworzy kategorie niepowiązane ze sobą żadnymi relacjami, natomiast drugi dostarcza układ hierarchiczny, ze wskazanymi relacjami między poszczególnymi kategoriami. Metody grupowania hierarchicznego można podzielić na nieostre, gromadzące oraz ostre, wyróżniające. W przypadku ostrych algorytmów grupowania każdy dokument przypisywany jest tylko i wyłącznie do jednej kategorii, natomiast algorytmy nieostre mogą przypisać dokument do kilku kategorii. Klasy powstałe w wyniku grupowania hierarchicznego można przedstawić w postaci dendrogramu (drzewa zależności)[40].

---

[37] Tamże, s. 117.

[38] Tamże, s. 67, 117-119.

[39] Tamże, s. 188-190.

[40] O grupowaniu dokumentów por. Ch. D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 349 i nast.; D. Jurafsky, J. H. Martin: dz. cyt., s. 679 i nast.

Charakterystyki wyszukiwawcze dokumentów przedstawione w postaci wektorów pozwalają przyporządkować konkretne dokumenty do zdefiniowanych, stałych klas na podstawie podobieństwa wektora dokumentu i wektora kategorii. Możliwe jest również automatyczne generowanie kategorii tematycznych na podstawie dodatkowej analizy podobieństwa pomiędzy dokumentami relewantnymi do zapytania. W procesie grupowania można wskazać dwa główne nurty: grupowanie oparte o wzorce oraz bezwzorcowe.

1.2.2.1. Grupowanie oparte o wzorce

Grupowanie oparte o wzorce polega na przypisaniu poszczególnych tekstów do jednej z ustalonych wcześniej klas tematycznych. Dla każdej kategorii tematycznej dokumentów należy wskazać wzorce, pozwalające przypisać do niej treść, czyli zdefiniować charakterystykę kategorii. Metoda ta świetnie sprawdza się np. w katalogach internetowych, które operują właśnie na ustalonych zbiorach tematycznych. Kategorie należy określać na tyle elastycznie, żeby można było przypisać do nich dokumenty tylko częściowo klasyfikujące się do danego zakresu tematycznego; ewentualnie można utworzyć dodatkową grupę INNE, chociaż nie jest to rozwiązanie eleganckie i nie w pełni profesjonalne. Grupa taka utrudnia zlokalizowanie dokumentu, ze względu na problemy z prawidłową kategoryzacją mogą zostać do niej przypisane dokumenty bardzo odległe od siebie tematycznie, co ma oczywiście negatywny wpływ na jakość wyszukiwania w danym systemie.

1.2.2.2. Grupowanie bezwzorcowe

Pewną alternatywą jest automatyczne tworzenie kategorii dostosowanych do posiadanej kolekcji dokumentów. Metoda ta nadaje się dobrze do zastosowania m.in. w wyszukiwarkach internetowych, ponieważ opiera się na zbiorze niesklasyfikowanych dokumentów. Dopiero na podstawie charakterystyk treściowych oraz rozkładu częstości podobnych reprezentacji system generuje grupy tematyczne i przypisuje do nich poszczególne dokumenty. Metoda ta zwana jest również taksonomią lub analizą skupień.

W przypadku zamkniętych, kontrolowanych systemów informacyjno-wyszukiwawczych dostępny zestaw słów kluczowych jednoznacznie określa zakres treściowy, ułatwiając użytkownikowi wybór najbardziej re-

lewantnego dokumentu. Z drugiej zaś strony, wyszukiwarki internetowe pracują w środowisku otwartym, bez obowiązujących powszechnie reguł tworzenia charakterystyk wyszukiwawczych. W związku z potrzebą standaryzowania i ujednolicenia trybu komunikacji wyszukiwarki z użytkownikiem w różnych wyszukiwarkach stosowano różne metody prezentacji treści dokumentów zgodnych z zapytaniem. Jednakże ze względów praktycznych, wynikających z komunikacyjnych nawyków użytkowników, najpopularniejszą metodą prezentowania tematyki dokumentu jest wyświetlenie kilku pierwszych zdań, bądź kilku zdań sąsiadujących z miejscem zlokalizowania w treści słowa kluczowego.

Zarówno w przypadku grupowania opartego o wzorce, jak i bezwzorcowego nazwy poszczególnych kategorii stanowią wyrażenia będące z definicji słowami kluczowymi charakteryzującymi w stopniu ogólnym treść dokumentów. Takie ogólne słowa kluczowe przypisane do dokumentów nie wpływają znacząco na zawężenie zbioru wyników będącego odpowiedzią systemu na zapytanie użytkownika, mogą jednakże ułatwić wskazanie dokumentów podobnych do siebie treściowo.

Niestety, pomimo dynamicznego rozwoju badań nad komputerowym przetwarzaniem języka naturalnego w Polsce, wydaje się, że nadal aktualne pozostają słowa M. Świdzińskiego: „Bardzo niepokojące jest zwłaszcza to, że lingwistyką informatyczną zajmują się w Polsce pojedynczy językoznawcy... Dużo więcej informatyków w Polsce pracuje w tej dziedzinie, niż lingwistów”[41]. Jako przyczynę takiego stanu rzeczy cytowany badacz podaje brak studiów lingwistycznych na poziomie uniwersyteckim, co przejawia się umiejscowieniem językoznawstwa na wydziałach filologicznych[42].

[41] Cyt. za: M. Świdziński: dz. cyt., s. 32.
[42] Tamże, s. 32.

Ustalenia terminologiczne oraz wybrane metody komputerowego przetwarzania języka naturalnego[43]

Książka niniejsza w dużym stopniu wykorzystuje statystyczne metody komputerowego przetwarzania tekstów, zarówno pojedynczych dokumentów, jak i, przede wszystkim, całych ich kolekcji. Gwoli uściślenia należy dodać, że na potrzeby książki analizowane są teksty z poszczególnych dokumentów, natomiast wnioski wyciągane są na podstawie porównania określonych właściwości dotyczących jednego dokumentu z wartościami tych samych cech stwierdzonych dla całego zbioru dokumentów. Wnioskowanie o prawidłowościach językowych przeprowadzane jest na podstawie statystycznej analizy odpowiednio dużych zbiorów tekstów.

Jak podaje Mieczysław Sobczak, **statystyka** jest nauką dotyczącą ilościowych metod badania **zjawisk** (inaczej procesów) **masowych**[44]. Pojęcie masowości zakłada badanie odpowiednio dużego **zbioru jednostek**, które cechują się podobnymi, ale nie identycznymi właściwościami. Wynikiem badań statystycznych są reguły bądź wnioski dotyczące uśrednionych wartości cech badanych zbiorowości. Te reguły to tzw. **prawidłowości statystyczne**. Badania statystyczne dotyczą **zbiorowości statystycznej** (populacji, masy statystycznej). **Populacja** oznacza zbiór elementów ob-

[43] Część niniejszego rozdziału została opublikowana w artykule P. Malak: *Metody statystyczne w komputerowym przetwarzaniu języka naturalnego*. „Toruńskie Studia Bibliologiczne" 2011, nr 1 (6), s. 49-62.

[44] Wprowadzenie do statystyki – por. M. Sobczyk: *Statystyka*. Wyd. 3 zmien., Warszawa: Wydaw. Naukowe PWN 2000, tegoż: *Statystyka. Podstawy teoretyczne przykłady – zadania*. Lublin: Wydaw. Uniwersytetu M. Curie-Skłodowskiej 1998.

jętych badaniem statystycznym. Poszczególne elementy składowe populacji nazywane są **jednostkami statystycznymi**, przy czym w obrębie jednej zbiorowości statystycznej można wyróżnić wiele jednostek statystycznych (np. podzbiór leksemów, zdań czy też całych tekstów badanego zbioru dokumentów)[45].

Autorzy pracy *Foundations of statistical natural language processing* w interesujący sposób streścili umiejscowienie i przynależność statystycznego nurtu NLP. Badania kwantytatywne nad językiem naturalnym zdefiniowali jako dyscyplinę łączącą wszystkie podejścia ilościowe do automatycznego przetwarzania języka, włączając modelowanie probabilistyczne, teorię informacji oraz algebrę liniową. Pomimo potencjalnej wieloznaczności tego terminu Manning i Schütze konkludują, że na przestrzeni ostatniej dekady **statystyczne NLP** było terminem używanym najpowszechniej do oznaczenia wszystkich prac nad przetwarzaniem języka naturalnego nie wprowadzających symboliki ani logiki[46].

Należy zgodzić się z powyższymi wywodami, ponieważ badania statystyczne języka naturalnego rzeczywiście korzystają z osiągnięć teorii informacji, teorii prawdopodobieństwa oraz rozwiązań algebry liniowej do przeprowadzenia wieloaspektowej analizy wyrażeń językowych. W takim też uniwersalnym znaczeniu będą używane w niniejszej książce terminy **lingwistyka kwantytatywna** czy też **lingwistyka statystyczna**.

Na opracowanie kwantytatywne zbioru dokumentów składają się w dużej części operacje mechaniczne przygotowujące poszczególne dokumenty do właściwego procesu analizy. Są to operacje takie jak np. wykluczenie z tekstu wyrazów znajdujących się na liście słów mało znaczących[47] (ang. *stop list*) w celu obniżenia kosztów przetwarzania elemen-

[45] Definicje poszczególnych terminów statystycznych – por. tegoż: *Statystyka*, dz. cyt., s. 11-13.

[46] Ch. D. Manning, H. Schütze: dz. cyt., s. XXXI-XXXII, stwierdzają: "A final remark is in order on the title we have chosen for this book. Calling the field Statistical Natural Language Processing might seem questionable to someone who takes their definition of a statistical method from a standard introduction to statistics. Statistical NLP as we define it comprises all quantitative approaches to automated language processing, including probabilistic modeling, information theory, and linear algebra. While probability theory is the foundation for formal statistical reasoning, we take the basic meaning of the term 'statistics' as being broader, encompassing all quantitative approaches to data (a definition which one can quickly confirm in almost any dictionary). Although there is thus some potential for ambiguity, Statistical NLP has been the most widely used term to refer to non symbolic and non logical work on NLP over the past decade, and we have decided to keep with this term".

[47] W informacji naukowej znany jest termin *wyrażenie nierelewantne (słowo nieznaczące)*, rozumiany jako wyrażenie jiw o małej wartości wyszukiwawczej. W niektórych jiw o nota-

tów tekstu, które nie wnoszą wartościowych informacji, zliczenie częstości wystąpień danego wyrazu (ang. *term frequency,* TF), czy porównanie liczby wystąpień poszczególnych wyrazów w różnych dokumentach badanego zbioru. Warto nadmienić, że w teorii informacji pod pojęciem *stop lista* funkcjonuje *słownik ujemny*. Jest to słownik otwarty, który zawiera wyrażenia języka naturalnego, nie mogące pełnić funkcji wyrażeń jiw, przy czym stop listy są najczęściej dodawane do języków słów kluczowych[48].

Wymienione wcześniej operacje tego typu, ważne dla dokonania poprawnej analizy dokumentu, nie wymagają udziału człowieka, mogą z powodzeniem zostać przeprowadzone przez specjalistyczne oprogramowanie. Zastosowanie komputerów do badań nad tekstami języka naturalnego pozwala na obniżenie kosztów operacji mechanicznych oraz zwielokrotnienie ilości tych operacji wykonanych w określonym czasie w porównaniu do analizy przeprowadzanej przez człowieka. W związku z tym, oczywistym jest fakt scedowania na komputery jak największej części prac związanych z opracowaniem zbioru dokumentów, pozostawiając człowiekowi kontrolę nad zautomatyzowanym procesem.

W bieżącym rozdziale zostaną zaprezentowane podstawy kwantytatywnej analizy tekstów języka naturalnego oraz wybrane metody komputerowego przetwarzania języka naturalnego, przydatne dla celów postawionych przed niniejszą książką. Przeprowadzona zostanie również dyskusja przyjętych w książce terminów.

2.1. Terminy przyjęte w książce

Na potrzeby niniejszej książki, łączącej elementy przetwarzanie języka naturalnego z praktycznymi aspektami informacji naukowej, przyjęto następujące, aplikacyjne definicje poszczególnych jednostek analizy tekstu.

Token – (inaczej **segment**), jest to najmniejsza, niepodzielna jednostka tekstu, której można przypisać interpretację morfosyntaktyczną (informację o części mowy i odpowiedniej kategorii morfosyntaktycznej). Tokeny są jednostkami uzyskiwanymi w wyniku bardzo szczegółowej analizy mor-

---

cji paranaturalnej, przeznaczonych do indeksowania swobodnego, wyrażenia takie podawane są w postaci listy lub poprzez określenie kategorii syntaktycznych, np. przyimki, zaimki, czy kategorii semantycznej, np. nazw dziedzin wiedzy. Por. hasło *wyrażenie nierelewantne* W: *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*, pod red. B. Bojar. Warszawa: SBP 2002, s. 302.

[48] Por. hasło *słownik ujemny*. W: *Słownik encyklopedyczny informacji...*, s. 245.

fosyntaktycznej tekstu. W wielu przypadkach **segment** jest tożsamy z **wyrazem słownikowym**, jednakże **wyraz** może często zostać podzielony na kilka **tokenów**, np. *byłbyś* = *był* + *by* + ś (gdzie token *-by* oznacza partykułę oznaczającą tryb warunkowy, natomiast *-ś* oznacza drugą osobę liczby pojedynczej). W niektórych z kolei sytuacjach **segment** jest dłuższy niż słowo gramatyczne, np. *po prostu*, *śmiać się*[49].

Należy przy okazji zaznaczyć, że w językach niefleksyjnych (jak np. język angielski), pojęcie **token** w zdecydowanej większości przypadków pokrywa się znaczeniowo z pojęciami **wyraz**, **słowo** oraz **hasło**. Stąd też, na potrzeby przetwarzania języka polskiego, trzeba było zweryfikować znaczenie tego terminu, wypracowane początkowo dla języka angielskiego. W pracach angielskich badaczy termin token bywa również stosowany dla oznaczenia różnych graficznie postaci tego samego słowa (np. *They* i *they*)[50].

Słowo – ciąg znaków pomiędzy delimitatorami tekstu, do których zostają zaliczone znaki przestankowe i spacje. Natomiast nie będą traktowane jako znaki separacji tekstu: łącznik (-) oraz apostrof (' – występujący pojedynczo, bez spacji lub innego znaku przestankowego w swoim bezpośrednim otoczeniu). W praktyce **słowo** oznacza fizyczną (graficzną) realizację danego leksemu[51].

Słowoforma – (inaczej **forma wyrazowa**), słowo o przypisanych cechach semantycznych i gramatycznych. W przypadku analiz kwantytatywnych tekstów pisanych języka naturalnego słowoforma tożsama jest z pojęciem **słowo**, ponieważ analizie statystycznej podlegają fizyczne elementy tekstów danego języka, czyli właśnie słowa[52].

---

[49] O pojęciu token (segment) por. m.in. A. Mykowiecka: *Inżynieria lingwistyczna...*, s. 65-66; M. Piasecki: *Cele i zadania...*, s. 260-264; A. Przepiórkowski: *Powierzchniowe przetwarzanie...*, s. 17-20. Bardzo szczegółową dyskusję procesu segmentacji można znaleźć w pracy tegoż: *Korpus IPI PAN. Wersja wstępna*. Warszawa: IPI PAN 2004, s. 13-15, 17-39.

[50] O angielskich definicjach terminu token por. m.in. Ch. D. Manning, H. Schütze: dz. cyt., s. 21-22; D. Jurafsky, J. H. Martin: dz. cyt., s. 193.

[51] Definicję terminu słowo w kontekście inżynierii lingwistycznej, w tym postulat traktowania apostrofu i dywizu jako integralnych elementów słowa por. m.in.: R. Hammerl, J. Sambor: *Statystyka dla...*, s. 17; tychże: *O statystycznych...*, s. 21; A. Przepiórkowski: *Korpus IPI PAN...*, s. 20; J. S. Bień: *Aparat pojęciowy...*, s. 24. Dyskusję apostrofu i łącznika jako elementów słów można znaleźć również w rozprawie doktorskiej M. Woliński: *Komputerowa weryfikacja gramatyki Świdzińskiego. Rozprawa doktorska przygotowana pod kierunkiem dr. hab. Janusza S. Bienia, prof. UW* [on-line], s. 50. [Dostęp: 17 września 2011]. Dostępny w World Wide Web: http://www.ipipan.eu/staff/m.wolinski/publ/mw-phd.pdf.

[52] Za: R. Hammerl, J. Sambor: *Statystyka dla...*, s. 18.

Leksem – zbiór słowoform, zawierający wszystkie poprawne formy gramatyczne danego słowa. Ze względu na operacje wstępnego formatowania i przygotowania tekstu do analiz leksem najczęściej wyrażany będzie w niniejszej książce w postaci **hasła**, czyli zwyczajowo przyjętej formy gramatycznej.

Hasło – (inaczej **wyraz słownikowy**, **forma podstawowa**, **lemat**) jedna, zwyczajowo przyjęta (kanoniczna) forma gramatyczna danego leksemu (np. bezokolicznik dla czasowników w języku polskim).

Termin – wyrażenie o ściśle ustalonym znaczeniu w danej dziedzinie nauki lub techniki. W takim znaczeniu terminy mogą być dobrymi **słowami kluczowymi** dla tekstów o określonej tematyce[53].

Wyraz – graficzna postać leksemu, hasła lub słowoformy, używane zamiennie w stosunku do owych trzech terminów. W bieżącej książce wyrazami będą ujednolicone postaci słów, uzyskane w wyniku operacji wstępnego przetwarzania tekstów analizowanych dokumentów[54].

Słowo kluczowe – wyrażenie z tekstu dokumentu lub zapytania informacyjnego charakteryzujące jego treść[55].

Próba – (**próba reprezentatywna**) część populacji generalnej pozwalająca na przeprowadzenie poprawnego procesu wnioskowania statystycznego, dotyczącego całej populacji. Próba jest zawsze skończona. W niniejszej książce jakikolwiek korpus, słownik jednojęzyczny czy zbiór tekstów o wspólnych cechach wyróżniających traktowany będzie jako próba z populacji generalnej, jaką jest nieskończony zbiór leksemów danego języka, stylu funkcjonalnego czy konkretnego autora[56].

Należy przy okazji odnotować pewne różnice terminologiczne pomiędzy językiem polskim a angielskim, które wynikają z różnic typów obu języków[57]. Nadmieniono już w niniejszej książce możliwość zastosowania angielskiego terminu **token** dla różnych graficznie postaci tego samego

---

[53] Zob. hasło *termin*. W: *Słownik encyklopedyczny informacji...*, s. 277.

[54] Za: tamże, s. 301 (hasło *wyraz*), różne kryteria definicyjne pojęcia wyraz dyskutuje również J. S. Bień: *Koncepcja słownikowej...*, s. 13-15.

[55] Por. hasło *słowo kluczowe*. W: *Słownik encyklopedyczny informacji ...*, s. 246.

[56] O pojęciu próba w statystycznych badaniach lingwistycznych por. m.in. R. Hammerl, J. Sambor: *Statystyka dla...*, s. 16-17.

[57] Odnotowanie różnic terminologicznych jest o tyle sensowne i usprawiedliwione, że same badania przetwarzania języka naturalnego zostały rozpoczęte w krajach anglosaskich (głównie USA), a poziom zaawansowania tych badań dla języka angielskiego jest najwyższy. Z powodu prymatu krajów anglojęzycznych w owych badaniach stosowana w nich terminologia jest oryginalnie pochodzenia angielskiego.

wyrazu. Termin *word token* w angielskiej literaturze przedmiotu stosowany jest na określenie każdego wystąpienia wyrazu w tekście (z uwzględnieniem powyższej uwagi). Z kolei dla oznaczenia różnych znaczeniowo słów stosowane jest pojęcie *word type*. W terminologii przyjętej na potrzeby niniejszej książki angielskiemu pojęciu token odpowiadają terminy słowo/wyraz, natomiast terminowi *word type* odpowiada hasło (wyraz słownikowy). Pewne wątpliwości znaczeniowe mogę pojawić się również dla pojęcia *term* (termin). Powszechnie przyjętym znaczeniem tego pojęcia w języku polskim jest wyrażenie o ściśle ustalonym znaczeniu w danej dziedzinie. Natomiast w tekstach anglojęzycznych poświęconych NLP określenie *term* wydaje się być stosowane zamiennie z określeniem *word type* dla oznaczenia każdego odmiennego znaczeniowo wystąpienia danego słowa. Bardzo często we wzorach związanych z przetwarzaniem tekstów języka naturalnego spotkać można oznaczenie **t**, pokrywające się znaczeniowo z pojęciem *word type*[58].

Jakiekolwiek wiążące i wiarygodne statystyki dotyczące tekstów języka naturalnego wymagają porównania danych dla wielu dokumentów. W analizie tekstów języka naturalnego najczęściej bada się prawidłowości związane z występowaniem poszczególnych wyrazów (zarówno w postaci słów, jak i haseł w rozumieniu przyjętych definicji). Zawartość słownictwa analizowanych dokumentów tworzy tzw. korpus badawczy. Liczba odnotowanych wyrazów, czyli pojemność badanego zbioru (inaczej wielkość korpusu), będącego podstawą do porównań, jest bardzo istotna w przypadku statycznych systemów informacyjno-wyszukiwawczych. Jeszcze istotniejsza jest ona w przypadku systemów dynamicznych, jakimi są np. wyszukiwarki internetowe[59]. W związku z rozmiarami zbiorów da-

[58] Por. m.in. cytowane już prace D. Jurafsky, J. H. Martin: dz. cyt., Ch. D. Manning, H. Schütze: dz. cyt., Ch. D. Manning, P. Raghavan, H. Schütze: dz. cyt., P. Jackson, I. Moulinier: dz. cyt.

[59] Pod pojęciem statycznych systemów informacyjno-wyszukiwawczych rozumiemy np. katalogi biblioteczne oraz inne źródła, których zawartość generowana jest przez wykwalifikowane w klasyfikowaniu treści osoby. Zawartość ta odzwierciedla treść opisanych dokumentów oraz pozwala zidentyfikować dokumenty odpowiadające zapytaniom użytkowników. Natomiast systemami dynamicznymi są m.in. wyszukiwarki internetowe indeksujące automatycznie ciągle zmieniającą się zawartość dokumentów, generowanych przez wielu użytkowników o różnych kwalifikacjach klasyfikacji treści. Opisy zindeksowanych dokumentów sprowadzane są zazwyczaj do wskazania słów kluczowych, nie ograniczonych jakimkolwiek słownikiem, odzwierciedlających z określonym prawdopodobieństwem zawartość treściową konkretnych dokumentów. Prawdopodobieństwo zgodności słowa kluczowego z treścią dokumentu wyznaczane jest na podstawie analizy frekwencyjnej słów występujących w tekście.

nych przetwarzanych w trakcie operacji NLP oczywista staje się konieczność dostarczenia tych danych w postaci jak najmniej angażującej dowolny system komputerowy. Jednym z powszechnie stosowanych sposobów jest optymalizacja dokumentu polegająca m.in. na pominięciu słów mało znaczących oraz sprowadzeniu poszczególnych wyrazów do postaci hasłowej. Poszczególne metody optymalizacji zostały przedstawione w dalszej części bieżącego rozdziału. Z kolei dla potrzeb efektywnego wyszukiwania informacji w dokumentach stosuje się optymalizację treści dokumentu do postaci **reprezentacji dokumentu**[60]. Opis odzwierciedlający treść dokumentu, ale niebędący dokładną kopią tekstu, stanowi reprezentację treści dokumentu i jest podstawą jakichkolwiek operacji związanych z analizą i porównywaniem treści dokumentów.

2.2. Analiza kwantytatywna tekstów

Analiza kwantytatywna języka naturalnego wykorzystuje bardzo duże zbiory danych do generowania wniosków o tekstach bądź języku. Metody statystyczne stosowane w badaniach NLP w określonym zakresie pozwalają uzyskać wiarygodne i wartościowe wyniki analiz przy niskich kosztach operacyjnych. Jak podaje np. A. Mykowiecka, analiza frekwencyjna znajduje zastosowanie w indeksowaniu lub klasyfikacji dokumentów, wskazywaniu kategorii tematycznej treści dokumentów lub określaniu języka tekstu. Oprócz pojedynczych elementów języka analizie mogą podlegać złączenia, czyli tzw. współwystępowanie składników. Określenie frekwencji występowania poszczególnych złączeń wyrazów może być wykorzystane np. przy wskazywaniu znaczenia wyrazów wieloznacznych (w zależności od częstości poszczególnych złączeń)[61].

Interesujące argumenty za stosowaniem metod ilościowych w nauce o języku ogólnie przedstawił A. Pawłowski, językoznawca specjalizujący się w lingwistyce korpusowej, w artykule *Empiryczne i ilościowe metody badań wobec naukowego statusu współczesnego językoznawstwa*. Spośród wielu podanych przez niego argumentów warto przytoczyć następujące:

---

[60] Dzięki reprezentacjom dokumentów, pełniącym funkcje charakterystyk wyszukiwawczych, można zwiększyć wydajność systemu (więcej dokumentów porównanych w danym czasie), obniżyć koszty wyszukiwania informacji oraz ułatwić użytkownikowi podjęcie decyzji o wyborze konkretnego dokumentu z listy proponowanych.

[61] O możliwościach zastosowania metod statystycznych do analizy języka naturalnego por. m.in. A. Mykowiecka: *Inżynieria lingwistyczna...*, s. 188-191.

- teksty i/lub system mogą być reprezentowane przez dane liczbowe, co z kolei pozwala na budowanie modeli matematycznych (funkcyjnych i probabilistycznych) w badaniach lingwistycznych,
- dane językowe na poziomie tekstu i systemu można poddać procesom segmentacji i kwantyfikacji, które to procesy są poznawczo uzasadnione,
- kwantyfikacji można poddać każdą cechę języka, w tym jego warstwę semantyczną[62].

Jak z kolei podaje M. Piasecki, analizując dany korpus tekstów można przyjąć jedną z dwóch podstawowych perspektyw badawczych:

- analizę pojedynczych wystąpień lematów,
- analizę danych statystycznych dotyczących dystrybucji lematów w całym korpusie[63].

Fundamentalne zasługi dla przybliżenia statystyki oraz możliwości jej zastosowań w badaniach nad językiem w Polsce wniosła prof. Jadwiga Sambor. Można tu wymienić m.in. jej publikację *Językoznawstwo statystyczne dla pracowników informacji naukowej* (Warszawa 1978) czy książki współautorskie Jadwigi Sambor i Rolfa Hammerla: *Statystyka dla językoznawców* (Warszawa 1990) oraz *O statystycznych prawach językowych* (Warszawa 1993)[64]. W swoich pracach prof. Sambor przedstawia wyczerpujący wstęp do rachunku prawdopodobieństwa i statystycznych metod analizy tekstu oraz podaje przykłady użycia poszczególnych opisywanych przez siebie metod. Bardzo dobry i wyczerpujący wstęp do rachunku prawdopodobieństwa, wnioskowania statystycznego oraz wykorzystania metod statystycznych w badaniach języka zawiera publikacja J. Sambor zatytułowana *Językoznawstwo statystyczne dla pracowników informacji naukowej,* Warszawa 1978. Zaś praca R. Hammerla i J. Sambor, *Statystyka*

---

[62] Za: A. Pawłowski: *Empiryczne i ilościowe metody badań wobec naukowego statusu współczesnego językoznawstwa*. W: *Metodologie językoznawstwa. Filozoficzne i empiryczne problemy w analizie języka*, pod red. P. Stalmaszczyka. Łódź: Wydawnictwo Uniwersytetu Łódzkiego 2010, s. 128.

[63] Za: M. Piasecki: *Automatyczne wydobywanie wiedzy o semantyce języka naturalnego z korpusu tekstów*. W: *Metodologie językoznawstwa. Filozoficzne...*, dz. cyt., s. 149. Przy czym pod pojęciem lematu kryje się grupa jednostek leksykalnych reprezentowanych przez dany leksem.

[64] Por. J. Sambor: *Językoznawstwo statystyczne dla pracowników informacji naukowej*, Warszawa: CINTE 1978; R. Hammerl, J. Sambor: *O statystycznych prawach językowych*. Warszawa: Zakład Semiotyki Logicznej Uniwersytetu Warszawskiego, Polskie Towarzystwo Semiotyczne 1993; tychże: *Statystyka dla językoznawców*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego 1990.

dla językoznawców, jest bardzo szczegółowym wprowadzeniem zarówno do metod statystycznych, teorii informacji, jak i praw oraz prawidłowości językowych uzyskanych na podstawie analiz statystycznych języka. Natomiast ci sami autorzy w pracy *O statystycznych prawach językowych* prezentują dość szczegółowo prawa oraz prawidłowości statystyczne dotyczące języka naturalnego wraz z dyskusją nad terminem **prawo językowe**.

Analizą statystyczną prawidłowości ilościowych w tekstach i języku zajmuje się **lingwistyka kwantytatywna**. Według definicji słownikowej prawidłowości te dotyczą przede wszystkim frekwencji (częstości) występowania wyrażeń i struktur językowych wszystkich poziomów języka. Ponadto w zakres lingwistycznych badań kwantytatywnych wchodzi prawdopodobieństwo występowania wyrażeń i struktur w różnych kontekstach, rodzajach tekstów czy stylach wypowiedzi. Analizowane są również zależności pomiędzy częstością występowania wyrażeń i struktur a innymi cechami tych wyrażeń i struktur lub ich wartością informacyjną (zależności te określa się mianem statystycznych praw językowych). Jak podają autorzy *Słownika encyklopedycznego terminologii...,* wyniki badań językoznawstwa statystycznego dowodzą, że częstość występowania wyrażeń jest ich cechą występującą systematycznie i jako taka powinna być uwzględniana w opisach systemów językowych, formalizacjach transformacji językowych, nauczaniu języków oraz innych pracach związanych z przetwarzaniem języka. Lingwistyka kwantytatywna traktowana jest w przywołanym *Słowniku* jako synonim terminu **lingwistyka statystyczna**[65].

Interesującą analizę oraz wprowadzenie do dyscypliny prezentuje również Adam Pawłowski w pracy *Metody kwantytatywne w sekwencyjnej analizie tekstu.* W publikacji tej znajdziemy zarówno dyskusję na temat przedmiotu, jak i celu lingwistyki kwantytatywnej oraz zwięzły, systematyczny opis poszczególnych praw i prawidłowości statystycznych dotyczących tekstów języka naturalnego. Autor zaprezentował także metody sekwencyjnego modelowania struktur tekstu oraz szczegółową dyskusję nad analizą sekwencyjną tekstów[66].

Metody kwantytatywne wykorzystywane są często w badaniach statystycznych nad dużymi zbiorami tekstów. Subdyscyplina zajmująca się badaniami tego typu nazywana jest **lingwistyką korpusową**. Przedstawiana

[65] Por. hasło *lingwistyka komputerowa*. W: *Słownik encyklopedyczny informacji...*, dz. cyt., s. 149.

[66] Por. A. Pawłowski: *Metody kwantytatywne w sekwencyjnej analizie tekstu*. Warszawa: KLF UW 2001, s. 6-74.

jest ona przez Barbarę Lewandowską-Tomaszczyk jako część językoznawstwa komputerowego, dla której materiałem badawczym są autentyczne teksty danego języka. Ta subdyscyplina bada usus językowy jako częstości występowania form (Lewandowska-Tomaszczyk używa tu terminu częstotliwość). Dalej twierdzi, że kognitywne językoznawstwo korpusowe łączy metody empiryczne (również korpusowe) z perspektywą kognitywną, która dąży do wykrycia związków między zdolnościami i strukturami poznawczymi a językowymi człowieka[67]. Z kolei M. Świdziński w pracy *Lingwistyka korpusowa w Polsce – źródła, stan, perspektywy* przybliża, w formie popularnonaukowej, historię rozwoju językoznawstwa jako dyscypliny naukowej, od strukturalizmu, poprzez generatywizm i lingwistykę formalną, aż po współczesną lingwistykę informatyczną. W pracy tej autor prezentuje również najważniejsze, według siebie, wyzwanie stojące przed lingwistyką informatyczną, czyli rozwiązywanie homonimii[68]. Z kolei w pracy *Dehomonimizacja i desynkretyzacja w procesie automatycznego przetwarzania wielkich korpusów tekstów polskich* autorzy podają, że **dehomonimizacją** (inaczej rozwiązywaniem homonimii) jest przypisanie odpowiedniej, pod względem morfologicznym, homoformy, czyli odkrywanie właściwej interpretacji wyrazu[69].

Należy również wspomnieć, że jeden rozdział cytowanej już pracy A. Mykowieckiej poświęcony jest statystycznym modelom języka. Autorka zaprezentowała w nim wprowadzenie do metod statystycznych w badaniach języka naturalnego oraz możliwości praktycznego zastosowania poszczególnych metod i technologii w opracowywaniu modeli języka naturalnego[70].

Warto przy okazji zwrócić uwagę na terminologię określającą frekwencję badanych jednostek językowych. W języku polskim występują dwa wyrażenia określające frekwencję, są to **częstość** oraz **częstotliwość**. W wielu publikacjach związanych z badaniami kwantytatywnymi nad tekstami języka naturalnego można spotkać się z pojęciem **częstotliwość występo-**

[67] Za: B. Lewandowska-Tomaszczyk: *Metody empiryczne i korpusowe w językoznawstwie kognitywnym*. W: *Metodologie językoznawstwa. Podstawy teoretyczne*, pod red. P. Stalmaszczyka. Łódź: Wydawnictwo Uniwersytetu Łódzkiego 2006, s. 262-264.

[68] Por. M. Świdziński: dz. cyt., s. 23-33.

[69] Por. M. Świdziński, M. Derwojedowa, M. Rudolf: *Dehomonimizacja i desynkretyzacja w procesie automatycznego przetwarzania wielkich korpusów tekstów polskich* [on-line]. [Dostęp: 10 stycznia 2012]. Dostępny w World Wide Web: http://www.mimuw.edu.pl/polszczyzna/PTJ/b/b58_187-199.pdf.

[70] Por. A. Mykowiecka: *Inżynieria lingwistyczna...*, s. 187-230.

wania, jednakże, zgodnie z definicją tego terminu, nie jest to zastosowanie poprawne. *Nowy słownik poprawnej polszczyzny* podaje następującą, dualną definicję hasła **częstotliwość**[71]:

- występowanie, powtarzanie się jakiegoś zjawiska, czynności w określonym czasie,
- liczba zdarzeń lub cykli zjawiska okresowego w jednostce czasu.

Wystąpienia poszczególnych jednostek w tekstach języka naturalnego nie są zjawiskiem okresowym, nie powtarzają się ze stałą regularnością ani w stałym przedziale czasowym. Dlatego na określenie liczby wystąpień elementów tekstu poprawnym wyrażeniem jest częstość. Zgodnie z definicją z przytoczonego *Nowego słownika*, **częstość** jest to „częste występowanie, powtarzanie się jakiejś czynności, zjawiska itp."[72]

W dalszej części definicji znajduje się co prawda odwołanie do znaczenia hasła **częstotliwość** w 2. znaczeniu, ale z dodatkowym określeniem *rzadko*[73].

Ze względu na niezależność wystąpień poszczególnych jednostek od regularnych, okresowych ram czasowych, na określenie liczby wystąpień słów, wyrazów oraz innych jednostek języka naturalnego w niniejszej książce stosowane będą wymiennie terminy **częstość** oraz **frekwencja**.

Badania statystyczne tekstów języka naturalnego mogą dotyczyć elementów różnych poziomów języka. W kolejnym podrozdziale zaprezentowano jednostki badań oraz definicje wybranych terminów stosowanych w niniejszej książce.

2.2.1. Jednostki badania kwantytatywnego tekstów

W lingwistyce kwantytatywnej jednostkami badań są podstawowe elementy różnych poziomów języka. Mogą to być np. elementy graficzne (litery, wyrazy tekstowe), fonologiczne (fonemy, sylaby), morfologiczne (morfemy, części mowy) czy składniowe (struktury zdaniowe i/lub frazowe). Jak podają J. Sambor i R. Hammerl, inwentarze tych jednostek w systemie językowym w aspekcie ściśle synchronicznym można traktować jako skończone i małe[74].

---

[71] Definicja cyt. za: *Nowy słownik poprawnej polszczyzny*, pod red. A. Markowskiego. Warszawa: Wydaw. Naukowe PWN 2002, s. 121.

[72] Cyt. za: tamże, s. 211.

[73] Tamże, s. 121.

[74] O jednostkach analizy na różnych poziomach języka por. m.in. R. Hammerl, J. Sambor: *Statystyka...*, dz. cyt., s. 16-17; tychże: *O statystycznych...*, s. 21-22.

Natomiast w przypadku analizy jednostek leksykalnych można przyjąć, że mamy do czynienia z populacjami nieskończonymi. A. Pawłowski w artykule *Uwagi na temat korpusu języka polskiego (reprezentatywność, aktualność, nazwa),* przy okazji dyskusji wokół metody reprezentacyjnej w badaniach języka, analizuje pojęcia skończoności oraz otwartości poszczególnych podsystemów systemu języka. Autor artykułu wyróżnia podsystemy zamknięte o niewielkiej i łatwej do określenia liczbie jednostek (np. system fonologiczny), systemy półotwarte, cechujące się przewagą liczbową jednostek potencjalnych nad jednostkami faktycznie obserwowanymi, przy czym dzięki zastosowaniu kombinatoryki można obliczyć liczbę jednostek potencjalnych (np. repertuar morfemów). Ostatni typ podzbiorów, wyróżniony przez referowanego badacza, stanowią systemy otwarte, czyli takie, w których liczba elementów jest teoretycznie skończona, lecz w praktyce nieprzeliczalna. Przykładem podsystemów otwartych jest system leksykalny danego języka[75].

Autorzy podręcznika *Statystyka dla językoznawców* wyróżniają, za Zygmuntem Salonim[76], następujące jednostki leksykalne:

- **słowo**,
- **słowoforma (forma wyrazowa)**,
- **leksem**,
- **wyraz**,
- **hasło**[77].

Jednakże przyjęte przez przywołanych autorów definicje owych terminów podane są w postaci skróconej, treściowo dostosowane do potrzeb komputerowego przetwarzania tekstów języka naturalnego. Ciekawą i obszerną dyskusję tych pojęć przeprowadził Janusz S. Bień, który w swoich pracach analizuje znaczenie i definicję terminów **wyraz**, **słowo** oraz **leksem**, a także wprowadza własną (obecnie dosyć powszechnie przyjętą) jednostkę – **fleksem**. Podkreślić jednak warto, że w dużym stopniu rozważania Bienia, bardzo wartościowe zarówno z punktu widzenia lingwistyki,

[75] Por. A. Pawłowski: *Uwagi na temat korpusu języka polskiego (reprezentatywność, aktualność, nazwa).* W: *Językoznawstwo w Polsce: stan i perspektywy,* pod red. S. Gajdy. Opole: PAN, Uniwersytet Opolski 2003, s. 165-166.

[76] Z. Saloni: *Kategoria rodzaju we współczesnym języku polskim.* W: *Kategorie gramatyczne grup imiennych*, pod red. R. Laskowskiego. Wrocław: Zakład Narodowy im. Ossolińskich 1976, s. 43-78, cyt za: S. Hammerl, J. Sambor: *Statystyka dla...*, s. 17.

[77] Por. S. Hammerl, J. Sambor: *O statystycznych...*, s. 17-19.

jak i przetwarzania tekstów języka naturalnego, wychodzą poza zakres tematyczny niniejszej książki[78].

Należy w tym miejscu również nadmienić, że wielu polskich badaczy analizujących komputerowo język naturalny sięga do opracowań Jana Tokarskiego, który w swych publikacjach rozważał możliwości zautomatyzowania niektórych etapów prac nad słownikami oraz wskazywał pomysły zrealizowania wybranych operacji automatycznie przez komputery. Interesujące rozważania nad znaczeniem terminu **wyraz** oraz **forma** można znaleźć w jego pracy *Fleksja polska*[79].

Definicje bardziej szczegółowe niż w pracach J. Sambor i R. Hammerla, a jednocześnie bliższe zastosowaniom w informacji naukowej, znajdziemy w przywoływanym już kilkakrotnie *Słowniku encyklopedycznym.* Omawiane tu terminy można zdefiniować za autorami *Słownika* następująco:

> **Wyraz** – traktowany jako synonim terminu **słowo**, jest wyrażeniem elementarnym. W językach naturalnych wyrazy składają się z morfemów leksykalnych lub z morfemów leksykalnych i gramatycznych. Termin wyraz może być interpretowany jako **leksem** (jednostka systemowa) albo jako **słowoforma**, czyli wyrażenie tekstowe. W celu ułatwienia jednoznacznego wskazania wyrazów w tekstach można dodatkowo zdefiniować je jako ciągi liter pomiędzy znakami delimitacji tekstu (spacje, znaki przestankowe). Ponadto pojedyncze wyrazy można określić jako ciąg morfemów, pomiędzy którymi nie może wystąpić żaden inny morfem[80].

W *Encyklopedii językoznawstwa ogólnego* termin wyraz definiowany jest (w rozumieniu potocznym) jako najmniejsza znacząca jednostka językowa, cechująca się względną samodzielnością składniową[81].

---

[78] O problemach definicyjnych terminów lingwistycznych oraz propozycjach ich rozwiązania por. J. S. Bień: *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji* [on-line]. [Dostęp: 14 września 2011]. Dostępny w World Wide Web: http://bc.klf.uw.edu.pl/12/2/emph.pdf; tegoż: *O pojęciu wyrazu morfologicznego* [on-line]. [Dostęp: 14 września 2011]. Dostępny w World Wide Web: http://bc.klf.uw.edu.pl/62/1/jsb-zsE.pdf; tegoż: *Aparat pojęciowy wybranych systemów przetwarzania tekstów polskich* [on-line]. [Dostęp: 14 września 2011]. Dostępny w World Wide Web: http://bc.klf.uw.edu.pl/84/1/JSB_BPTJ_LXII.pdf. Rozważania te są rozwinięciami ustaleń wprowadzonych przez J. Tokarskiego.

[79] Dyskusję znaczenia terminów wyraz oraz forma por. J. Tokarski: *Fleksja polska*. Wydanie III z uzupełnieniami. Warszawa: Wydawnictwo Naukowe PWN 2001, s. 20-24.

[80] Por. hasło *wyraz*. W: *Słownik encyklopedyczny informacji...*, s. 301.

[81] Por. hasło *wyraz*. W: *Encyklopedia językoznawstwa...*, s. 595.

Słowoforma – jest wyrażeniem będącym elementem tekstu. Stanowi realizację leksemu poprzez nadanie mu odpowiedniej formy językowej oraz połączenie z odpowiednim **morfemem**[82].

Morfem – jest to najmniejsze wyrażenie przekazujące znaczenie. Można wyróżnić **morfemy gramatyczne** (fleksyjne oraz słowotwórcze) oraz **morfemy leksykalne** (rdzenie)[83].

Termin – jest wyrażeniem o ściśle ustalonym znaczeniu w danej dziedzinie nauki lub techniki[84].

Termin **leksem** nie został zdefiniowany w *Słowniku encyklopedycznym* bezpośrednio. Z definicji terminu **wyraz** można wywnioskować, że leksem jest to wyrażenie poziomu leksykalnego, czyli wyraz systemowy[85].

Ponadto *Słownik encyklopedyczny* podaje trzy inne definicje, przydatne dla niniejszej książki. Są to pojęcia: **słowa klucze**, **słowo kluczowe** oraz **temat**.

Słowa klucze – są to wyrazy cechujące się w danym tekście lub korpusie tekstów frekwencją znacząco większą niż w danym języku naturalnym. Stanowią one wykładniki głównych tematów tekstu, są również charakterystyczne dla danego autora[86].

Słowo kluczowe – jest to wyrażenie z tekstu dokumentu lub zapytania informacyjnego charakteryzujące jego treść. W przypadku dokumentów słowa kluczowe pochodzą często z tytułu lub tytułów rozdziałów[87].

Z kolei w pracy *Języki informacyjno-wyszukiwawcze. Katalogi rzeczowe,* pojęcie słowa kluczowego zostało zaprezentowane w dwóch znaczeniach: jako wyrażenie charakteryzujące treść dokumentu, przejmowane z tekstu dokumentu oraz jako jednostka leksykalna języka słów kluczowych[88]. W niniejszej książce termin **słowo kluczowe** stosowany jest w znaczeniu wyrażenia reprezentującego treść dokumentu.

---

[82] Por. hasło *słowoforma*. W: *Słownik encyklopedyczny informacji...*, s. 246.

[83] Por. hasło *morfem*. W: Tamże, s. 163. Obszerną dyskusję znaczenia terminu oraz typów morfemów, z uwzględnieniem aspektu języków informacyjno-wyszukiwawczych, można również znaleźć w B. Bojar: *Językoznawstwo dla studentów informacji naukowej*. Warszawa: Wydawnictwo SBP 2005, s. 117-119.

[84] Por. hasło *termin*. W: *Słownik encyklopedyczny informacji...*, s. 277.

[85] Na podstawie hasła *wyraz*. W: Tamże, s. 301.

[86] Por. hasło *słowa klucze*. W: Tamże, s. 242.

[87] Por. hasło *słowo kluczowe*. W: Tamże, s. 246.

[88] Za: J. Sadowska, T. Turowska: *Języki informacyjno-wyszukiwawcze. Katalogi rzeczowe*. Warszawa: CUKB SBP 1990, s. 146.

Przy okazji dyskusji nad definicjami terminu słowo kluczowe nie można pominąć książki Wiesława Babika *Słowa kluczowe*. Stanowi ona najbardziej kompletne z polskich opracowań dotyczących procesów wyszukiwania informacji z zastosowaniem słów kluczowych. W systematyczny sposób prezentuje funkcje słów kluczowych, głównie w ujęciu jednostek języka informacyjno-wyszukiwawczego, w procesach indeksowania i wyszukiwania informacji. Analiza została przeprowadzona w szerokim spektrum wykorzystania słów kluczowych w różnych dziedzinach, z uwzględnieniem aktualnych ustaleń kognitywistycznych. Jak słusznie zauważa jej autor, pomimo rosnącego znaczenia słów kluczowych w wyszukiwaniu informacji, wiele problemów jest rozpoznanych w niewystarczającym stopniu[89].

Temat – definiowany również jako przedmiot dokumentu, to, czego dotyczą zawarte w dokumencie informacje. W informacji naukowej utożsamiany niekiedy z głównym przedmiotem dokumentu, znaczeniowo najważniejszym, dla omówienia którego powstał dokument[90].

Natomiast termin **hasło** został w *Słowniku encyklopedycznym* zdefiniowany wyłącznie w kontekście zastosowania w systemach informacyjno-wyszukiwawczych jako wyrażenie o funkcji porządkującej lub wyszukiwawczej w danym zbiorze informacyjnym (słownik, indeks, tekst, zbiór charakterystyk wyszukiwawczych dokumentów)[91].

Nieco odmiennie definiują wymienione pojęcia autorzy prowadzący badania w zakresie komputerowego przetwarzania języka naturalnego. Największe różnice dotyczą terminów **słowo**, **wyraz** oraz **hasło**. Definicja słownikowa utożsamia ze sobą dwa terminy: **słowo** oraz **wyraz**. Natomiast w pracach lingwistycznych spotykamy wyraźne zróżnicowanie znaczeń przypisywanych obu pojęciom. J. Sambor definiuje **słowo** jako jednostkę tekstu (lub języka) wyodrębnianą w procedurze segmentacyjnej, odpowiadającą w większości przypadków ciągowi liter pomiędzy odstępami. Z kolei termin **wyraz** wspomniana badaczka traktuje jako pojęcie nadrzędne do terminów słowo, słowoforma i leksem. W jej pracach termin **wyraz** używany jest zamiast wskazanych trzech terminów, w kontekście wskazującym jednoznacznie rodzaj zastępowanej jednostki[92].

---

[89] W. Babik: *Słowa kluczowe*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego 2010, s. 9.

[90] Por. hasło *temat*. W: *Słownik encyklopedyczny informacji...*, s. 272.

[91] Por. hasło *hasło*. W: Tamże, s. 76.

[92] Definicje terminu *słowo* oraz *wyraz* por. R. Hammerl, J. Sambor: *Statystyka dla...*, s. 17-19; tychże: *O statystycznych...*, s. 21-22. Definicję techniczną terminu *słowo*, jako ciągu znaków pomiędzy dwiema spacjami, przyjmuje również m.in. A. Mykowiecka; por. A. Mykowiecka: *Inżynieria lingwistyczna...*, s. 67.

Termin **hasło** w pracach dotyczących przetwarzania języka naturalnego definiowany jest jako zwyczajowo przyjęta w leksykografii danego języka forma gramatyczna leksemu (np. bezokolicznik dla czasowników w języku polskim). Pojęcie hasła można zdefiniować również jako zbiór słowoform reprezentowany przez określoną postać danej słowoformy[93].

Przytoczone powyżej terminy są podstawowymi pojęciami stosowanymi w przetwarzaniu tekstów. Określają one m.in. jednostki badania statystycznego wyrażeń języka naturalnego. Jednostki te cechują się konkretnymi cechami statystycznymi, które zostaną zaprezentowane poniżej. Natomiast definicje przyjęte na potrzeby niniejszej książki zostały zdefiniowane w dalszej części bieżącego rozdziału, łącznie z pozostałymi wykorzystywanymi definicjami.

2.2.2. Cechy statystyczne jednostek leksykalnych

Jednostki tekstu lub języka w danej zbiorowości statystycznej mogą być badane ilościowo ze względu na określoną cechę statystyczną X. Różne realizacje liczbowe x_i tej cechy w przypadku poszczególnych badanych jednostek odwzorowują ich zróżnicowanie pod kątem danej cechy X. Cechy statystyczne, ze względu na sposób ich zróżnicowania, można podzielić na:

- **cechy ilościowe**, które z kolei można podzielić na **ciągłe** (mierzalne – w danym przedziale wartości zmienne mogą przyjmować dowolne wartości liczbowe) lub **skokowe** (przeliczalne – w danym przedziale wartości zmienne mogą przyjmować tylko określone wartości liczbowe, np. liczby naturalne); w badaniach lingwistycznych częściej analizuje się cechy przeliczalne,
- **cechy jakościowe**, które nie są wyrażane liczbami, np. rodzimość lub obcość leksemu, złożoność lub prostota zdania[94].

Podstawową kategorią stosowaną w ilościowych obliczeniach statystycznych jest **częstość absolutna (frekwencja) F**. Jest to wskaźnik liczbowy otrzymany drogą sumowania jednostek wchodzących w skład danej próby. Podstawą sumowania mogą być wystąpienia poszczególnych jednostek bądź też wartości konkretnej cechy określającej dane jednostki. Częstość występowania słów w tekście jest cechą ilościową przeliczalną, o wartościach wyrażanych za pomocą liczb naturalnych.

---

[93] Por. m.in. R. Hammerl, J. Sambor: *Statystyka dla...*, s. 18; tychże: *O statystycznych...*, s. 21.
[94] Por. R. Hammerl, J. Sambor: *Statystyka dla...*, s. 19; M. Sobczyk: *Statystyka*, s. 12-13, 92-113.

Częstość absolutną **F** można przedstawić za pomocą następującego wzoru:

$$F = \sum_{i=1}^{n} f_i$$

gdzie:

F – częstość absolutna,

n – liczebność zbioru analizowanych dokumentów,

$\mathbf{f_i}$ – częstość wystąpienia danego słowa w kolejnym dokumencie[95].

Ponieważ jednym z elementów analizy tekstów na potrzeby niniejszej książki jest zbadanie możliwości wskazania słów kluczowych m.in. na podstawie częstości wystąpień, dla każdego wyróżnionego słowa wskazywana będzie również częstość średnia **f**, określana wzorem:

$$\bar{f} = \frac{F}{n} = \frac{(\Sigma_{i=1}^{n} f_i)}{n}$$

gdzie:

F – częstość absolutna,

n – liczebność zbioru analizowanych dokumentów,

$\mathbf{f_i}$ – częstość wystąpienia danego słowa w kolejnym dokumencie[96].

Na potrzeby badań opisanych w niniejszej książce jednostki analizowanych tekstów zostaną przeliczone w celu uzyskania wskaźnika ich frekwencji odpowiednio w poszczególnych dokumentach oraz w całości korpusu badawczego. Podejście takie jest elementem jednej z podstawowych technik statystycznej analizy tekstów, o czym traktuje dalsza część niniejszego rozdziału (poświęcona wybranym metodom i technikom NLP).

Dla korpusów zróżnicowanych wewnętrznie podaje się wskaźniki określające zróżnicowanie frekwencji danej jednostki w poszczególnych częściach korpusu. Podstawowym wskaźnikiem równomierności rozkładu jest dyspersja. Dyspersja (rozrzut) danej cechy mierzalnej opisuje zróżnicowanie jednostek badanego zbioru ze względu na tę cechę. Podstawowe miary dyspersji odzwierciedlają rozrzut wartości danej cechy wokół średniej arytmetycznej w badanym zbiorze. Jedną ze stosowanych w statystyce miar zmienności jest odchylenie standardowe **s**, które okre-

[95] Wzór na częstość absolutną słów w tekście por. I. Kurcz i in.: *Słownik frekwencyjny polszczyzny współczesnej*. Kraków: Instytut Języka Polskiego PAN 1990, s. L.

[96] Wzór na częstość średnią por. tamże, s. L.

śla przeciętne odchylenie częstości danej jednostki od częstości średniej dla całego zbioru.

Odchylenie standardowe określane jest wzorem[97]:

$$s = \sqrt{\frac{\Sigma_{i=1}^{n} (f_i - \bar{f})^2}{n - 1}}$$

Kolejną miarą jest współczynnik zmienności **v** określający relatywne odchylenie frekwencji danego elementu od częstości średniej[98].

$$v = \frac{s}{\bar{f}}$$

Jednakże, jak podają autorzy *Słownika frekwencyjnego,* miary te są w zbyt dużym stopniu zależne od wartości średniej, w związku z czym na potrzeby własnych badań wprowadzili wskaźnik **dyspersji złożonej**. Dyspersja złożona **D**, dostosowana do korpusu tekstów, wyrażana jest wzorem[99]:

$$D = 100 \text{ x } (1 - \frac{v}{\sqrt{n}}) = 100 \text{ x } (1 - \frac{\sqrt{\Sigma_{i=1}^{n} (f_i - \bar{f})^2}}{\sqrt{n (n - 1)} \text{ x } \bar{f}})$$

Analiza statystyczna elementów z określoną cechą statystyczną w systemie pozwala ustalić tzw. **udział** elementów w systemie, znany również jako **częstość relacyjna U**. Częstość relacyjna wyrażona jest następującym wzorem[100]:

$$U = \frac{F \text{ x } D}{100}$$

Autorzy *Statystyki* jako przykłady udziałów podają m.in. udział słownictwa częstego lub rzadkiego w tekście, czy też udział słownictwa rodzimego i obcego w określonym słowniku danego języka[101].

---

[97] Wzór na odchylenie standardowe por. tamże, s. L.

[98] Wzór na współczynnik zmienności por. tamże, s. L.

[99] O dyspersji słownictwa por. tamże, s. LI; R. Hammerl, J. Sambor: *Statystyka dla...*, s. 50 i nast.; J. Sambor: *Językoznawstwo statystyczne dla pracowników informacji naukowej*. Warszawa 1978, s. 51-53. O miarach zmienności w statystyce por. też M. Sobczyk: *Statystyka*, s. 45-53.

[100] Por. I. Kurcz i in.: *Słownik frekwencyjny...*, s. LI.

[101] O określaniu wartości średnich oraz ich odchyleń w badaniach językoznawczych por. R. Hammerl, J. Sambor: *Statystyka dla...*, s. 44-72.

2.2.3. Zależności leksykalne

Jak wynika z badań przeprowadzanych przy tworzeniu słowników frekwencyjnych, słownictwo języka naturalnego można podzielić z pewnym przybliżeniem na cztery strefy leksyki. Przynależność do konkretnej strefy zależy od wzajemnego stosunku wskaźników **F**, **D** i **U** określających częstość występowania poszczególnych słów w tekście. Autorzy *Słownika frekwencyjnego współczesnej polszczyzny* wyróżnili następujące strefy:

- strefa **słownictwa gramatycznego** – należą do niej hasła o największych częstościach, są to: podstawowe proste przyimki (takie jak: w, z, za), spójniki (że, i), zaimki (on) oraz czasowniki posiłkowe (być). Technicznie do strefy tej zaliczane są słowa o bardzo wysokich wartościach częstości absolutnej F i częstości relacyjnej U oraz o wartościach dyspersji złożonej D > 80[102],
- strefa **słownictwa podstawowego** – należą do niej hasła o bardzo wysokiej częstości, jak: rzeczowniki (człowiek, związek, część), przymiotniki (wielki, główny, inny), czasowniki, przysłówki i in. Technicznie zalicza się tu słowa o wysokich częstościach oraz równomierności rozkładu częstości D > 50[103],
- strefa **słownictwa charakterystycznego** – znajdują się tutaj hasła częste, ale ograniczone swym występowaniem do określonego typu publikacji (np. terminy techniczne, chemiczne, matematyczne, socjologiczne itp.). Technicznie, słownictwo charakterystyczne cechuje się wysoką częstością absolutną F oraz bardzo niską częstością relacyjną U, bliską zera. W *Słowniku frekwencyjnym* kryterium przynależności do strefy charakterystycznej wyrażone jest zależnościami: $F \geq 10$, $D \leq 50$[104],
- strefa **słownictwa rzadkiego** – zalicza się tu słownictwo, którego frekwencje mieszczą się poniżej średniej[105].

W pracy poświęconej różnicom leksykalnym pomiędzy stylami funkcjonalnymi polszczyzny pisanej, Irena Kamińska-Szmaj podała wyznaczniki liczbowe przynależności danego leksemu do konkretnej grupy. Zanalizowała statystyczne charakterystyki słownictwa pięciu stylów funkcjonalnych na podstawie danych ze *Słownika frekwencyjnego*. Ze względów na częstości wystąpień podzieliła słownictwo na trzy klasy: 1. Bardzo częste,

[102] Por. tamże, s. LV.
[103] Por. tamże, s. LVI.
[104] Tamże, s. LVI.
[105] O strefach leksyki por. m.in. I. Kurcz et.al.: *Słownik frekwencyjny...*, s. LVLVI.

2. Częste i średnio częste (połączone słownictwo podstawowe i gramatyczne) 3. Rzadkie. Do słownictwa bardzo częstego zalicza wyrazy o częstościach $f \geq 100$, do klasy wyrazów częstych i średnio częstych wyrazy o częstościach z przedziału $10 \leq f < 100$, zaś do słownictwa rzadkiego jednostki leksykalne o frekwencjach $f < 10$, która to wartość stanowi średnią częstość występowania wyrazów w tekstach[106].

W tabeli 2. zaprezentowana jest lista 20 leksemów języka polskiego o najwyższych frekwencjach, notowanych w *Słowniku frekwencyjnym*[107].

Tabela 2. Lista 20 leksemów najczęściej występujących w języku polskim.

Wyraz	Częstotść F	Procentowy udział w tekście	Wyraz	Częstość F	Procentowy udział w tekście
w	16316	3,26	**to**	5113	1,02
i	12385	2,47	**że**	4314	0,86
być	9621	1,92	**a**	3226	0,64
się	9302	1,86	**o**	3072	0,61
na	8600	1,72	**ja**	3028	0,60
nie	8341	1,66	**który**	2997	0,59
z	8310	1,66	**mieć**	2591	0,51
on	6650	1,33	**jak**	2251	0,45
do	5854	1,17	**co**	2224	0,44
ten	5743	1,14	**ale**	1963	0,39

Źródło: Opracowanie własne na podstawie danych ze *Słownika frekwencyjnego*.

Na liście tej leksem, który można zaliczyć do strefy słownictwa charakterystycznego (a jest nim wyraz *produkcja)* znajduje się dopiero na 138 pozycji (z częstością równą 363)[108]. Poza nim około 320 pierwszych pozycji zajętych jest przez słownictwo gramatyczne i podstawowe. Wyrazy

[106] Por. I. Kamińska-Szmaj: *Różnice leksykalne między stylami funkcjonalnymi polszczyzny pisanej. Analiza statystyczna na materiale słownika frekwencyjnego*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego 1990, s. 19-40.

[107] Częstości obliczane były dla korpusu 500 000 słów składającego się z pięciu równych części wyznaczanych przez styl funkcjonalny analizowanego piśmiennictwa (teksty popularnonaukowe, drobne wiadomości prasowe, publicystyka, proza artystyczna, dramat artystyczny). Por. I. Kurcz i in.: dz. cyt., s. XI.

[108] Por. I. Kurcz i in.: *Słownik frekwencyjny*, s. 802.

specjalistyczne są nierównomiernie rozmieszczone po całej objętości listy frekwencyjnej. Około 70% słownictwa naturalnego języka polskiego, odnotowanego w *Słowniku frekwencyjnym* wykazuje się częstościami poniżej 10 wystąpień[109].

Przytaczany *Słownik frekwencyjny* oraz publikacje oparte na podanych w nim wynikach badań, są najbardziej obszernym i kompletnym opracowaniem danych statystycznych dotyczących tekstów języka polskiego. Jednakże należy mieć świadomość, że język naturalny nie jest tworem statycznym, niezmiennym. Jest on nie tylko wytworem intelektualnym, ale także podstawowym narzędziem komunikacji i samookreślania własnych relacji za światem zewnętrznym. Jest to struktura bardzo elastyczna i dynamiczna, wykorzystywany zasób słownictwa zmienia się równolegle ze zmianami wprowadzanymi przez człowieka w świecie zewnętrznym. W związku z tą zmiennością niektóre cechy statystyczne odnotowane przez twórców *Słownika frekwencyjnego* nie są obecne we współczesnym języku. Przypomnijmy, że korpus dla przytaczanych badań stanowiły teksty języka polskiego z lat sześćdziesiątych XX wieku. Współczesne słownictwo odzwierciedla zmiany zarówno w zakresie technologii, jak i komunikacji społecznej oraz zmiany społeczno-socjalne, jakie zaszły przez pięćdziesiąt lat, które upłynęły od czasów utworzenia owego korpusu.

Obecnie najbardziej aktualne dane dotyczące charakterystyk statystycznych współczesnego języka polskiego zawiera *Korpus języka polskiego Wydawnictwa Naukowego PWN*, w ramach którego udostępnione jest m.in. zestawienie frekwencyjne. Pełna wersja korpusu, zrównoważona gatunkowo, składa się z 2070 próbek tekstów pochodzących z książek beletrystycznych i niebeletrystycznych, prasy, stron internetowych, ale także z druków ulotnych oraz rozmów. Uwzględnione przez PWN teksty pochodzą z lat 1920-2005, przy czym z okresu 1990-2005 pochodzi 78% słownictwa ujętego w korpusie. Łącznie zbiór tekstów liczy ponad 2 2000 000 słów tekstowych (około 4 000 000 słów po uwzględnieniu artykułów prasowych z dziennika *Rzeczpospolita*)[110].

Udostępniona nieodpłatnie na stronach PWN *Lista słów* rejestruje 3 707 391 wyrazów, sprowadzonych następnie do 246 697 form podsta-

[109] Na podstawie danych z tamże.

[110] Cyt. za: *Proporcje w pełnej wersji sieciowej korpusu* [on-line]. [Dostęp: 11 września 2011]. Dostępny w World Wide Web: http://korpus.pwn.pl/strukt_full.php.

wowych. Autorzy listy frekwencyjnej PWN podają, że w przypadku homonimów korzystano z proporcji ich znaczeń wskazanych w *Słowniku frekwencyjnym*. Autorzy opisywanej listy konkludują, że dla niektórych leksemów w tekstach współczesnych rozkład częstości jest taki sam, jak w tekstach analizowanych na potrzeby *Słownika frekwencyjnego*, zaś w przypadku słów nowych lub tych, których frekwencje we współczesnym języku wzrosły lub zmalały w porównaniu do języka z lat 60. XX wieku, uwidacznia się różnica proporcji występowania poszczególnych znaczeń[111].

Jednakże, inaczej niż *Słownik frekwencyjny,* lista słów PWN nie podaje danych statystycznych dla poszczególnych stylów funkcjonalnych piśmienniczo. W tabeli 3. zaprezentowano porównanie liczby wystąpień 20 najczęstszych leksemów z obu list frekwencyjnych.

Tabela 3. Porównanie frekwencji dwudziestu najczęściej występujących leksemów w korpusie polszczyzny z lat 60. XX wieku oraz w korpusie języka polskiego Wydawnictwa Naukowego PWN z lat 1920-2005.

Słownik frekwencyjny polszczyzny współczesnej			Lista słów PWN		
leksem	**częstość F**	**udział procentowy w korpusie**	**leksem**	**częstość F**	**udział procentowy w korpusie**
w	16316	3,26	**w**	120000	3,24
i	12385	2,47	**i**	95966	2,59
być	9621	1,92	**być**	78003	2,10
się	9302	1,86	**się**	76694	2,07
na	8600	1,72	**z**	65960	1,78
nie	8341	1,66	**na**	62634	1,69
z	8310	1,66	**nie**	56899	1,53
on	6650	1,33	**on**	53257	1,44
do	5854	1,17	**do**	43683	1,18
ten	5743	1,14	**ten**	41122	1,11
to	5113	1,02	**to**	37244	1,00
że	4314	0,86	**że**	31640	0,85

[111] Jako przykład podany został leksem **dział**, za: tamże.

Słownik frekwencyjny polszczyzny współczesnej			Lista słów PWN		
leksem	**częstość F**	**udział procentowy w korpusie**	**leksem**	**częstość F**	**udział procentowy w korpusie**
a	3226	0,64	**a**	28656	0,77
o	3072	0,61	**który**	24090	0,65
ja	3028	0,60	**o**	23882	0,64
który	2997	0,59	**mieć**	21693	0,59
mieć	2591	0,51	**jak**	18572	0,50
jak	2251	0,45	**tak**	16497	0,44
co	2224	0,44	**ja**	16338	0,44
ale	1963	0,39	**co**	16160	0,44

Źródło: Opracowanie własne na podstawie danych frekwencyjnych z obu omawianych korpusów.

Można zauważyć nieznaczne przesunięcia w zakresie częstości poszczególnych leksemów, ale nie są to wartości znaczące. Wśród 20 najczęstszych lematów jedynie *tak* pojawił się jako nowy element w porównaniu do polszczyzny lat 60. ubiegłego wieku.

Analizując wyniki badań frekwencyjnych dla słownictwa w językach naturalnych można zauważyć tendencję nieprzenikania się wyrazów z różnych stref. W tekstach danego języka o tematyce ogólnej łatwo na podstawie analizy frekwencji wskazać strefę, do której należy dany wyraz. Sytuacja ta ulega zmianie w przypadku tekstów specjalistycznych. Oczywisty jest fakt, że w zależności od dziedziny używane jest inne słownictwo specjalistyczne. Natomiast dzięki ustaleniu tego słownictwa dla poszczególnych dziedzin możliwe staje się wyznaczenie słów, które wpływają na wartość informacyjną dokumentów poświęconych zagadnieniom z danej dziedziny. Wyrazy charakterystyczne o dużych częstościach mogą być wykorzystywane jako słowa kluczowe w indeksowaniu automatycznym.

2.3. Wybrane metody reprezentacji treści dokumentów

W celu zwiększenia efektywności wyszukiwania dokumentów w systemach informacyjno-wyszukiwawczych stosuje się ustalone metody przybliżania treści dokumentów za pomocą krótkich tekstów: streszczeń, zestawów słów kluczowych, deskryptorów lub haseł przedmiotowych ze słownika. Takie przybliżenie nazywane jest reprezentacją treści dokumentu.

Wybrana, konkretna metoda reprezentowania treści dokumentu utożsamiana jest często z **modelem języka**. Pod pojęciem tym rozumiana jest funkcja przypisująca miary prawdopodobieństwa wystąpienia do poszczególnych ciągów znaków z zadanego zbioru słownictwa. Zdefiniowaną w powyższy sposób funkcję dla modelu języka **M** nad alfabetem **Σ** zapisuje się następująco[112]:

$$\sum_{s \in A} P(s) = 1$$

gdzie:

s – wyraz należący do alfabetu A,

A – zadany alfabet,

P(s) – prawdopodobieństwo wystąpienia wyrazu **s**[113].

Jest to model statystyczny, oparty na tekstach. Podobną definicję **modelu języka** przyjmują autorzy pracy *Natural Language Processing,* określając dany model jako dystrybucję prawdopodobieństwa reprezentującą uwarunkowania statystyczne rządzące tworzeniem zapytań. Tak pojęte modelowanie języka traktowane jest jako pewna alternatywa dla podejścia klasycznego, gdyż koncentruje się na generowaniu zapytań dotyczących poszczególnych dokumentów, a nie całej ich kolekcji, przy czym tworzenie zapytań traktowane jest jako proces losowy[114].

Z kolei D. Jurafsky i J. H. Martin podają, że pojęcie modelu języka w badaniach związanych z rozpoznawaniem mowy rozumiane jest tradycyjnie jako statystyczny model ciągów wyrazów i traktują je zamiennie z pojęciem gramatyka, używając obu terminów w zależności od kontekstu[115]. Tak zdefiniowany model należy kojarzyć z N. Chomsky'im. Jest to mechanizm rozpoznający i generujący zdania gramatyczne.

---

[112] Por. Ch. D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 238.

[113] Funkcja modelu języka por. tamże, s. 238.

[114] Por. P. Jackson, I. Moulinier: dz. cyt., s. 43.

[115] Za: D. Jurafsky i J. H. Martin: dz. cyt., s. 191.

Dla badań dotyczących zagadnień zaawansowanych, jak np. rozpoznawanie mowy, sprawdzanie pisowni czy automatyczne tłumaczenie, stosuje się kompleksowe modele języka (np. bezkontekstowe gramatyki probabilistyczne). Natomiast na potrzeby wyszukiwania dokumentów w zbiorze najczęściej korzysta się z **modeli unigramowych** (inaczej **prostych**, ang. *unigram language models*). Modele takie są zazwyczaj wystarczające do wskazania tematu analizowanego tekstu. Ponadto, unigramy są łatwiejsze i tańsze do zaimplementowania oraz zastosowania w celu wyszukiwania informacji niż złożone modele języka[116].

Według Piotra Gawrysiaka reprezentacja unigramowa (ang. *unigram model*) jest najprostszą i powszechnie stosowaną metodą reprezentacji treści dokumentu. Unigram jest wektorem, który rejestruje fakt wystąpienia lub braku danego słowa w treści dokumentu (reprezentacja binarna) lub częstości wystąpień poszczególnych wyrazów (reprezentacja częstościowa). W celu uniezależnienia wartości wektorów częstościowych można zamiast częstości bezwzględnej obliczyć częstość względną słów w dokumentach w kolekcji[117]. Dodatkowo, za McCallumem i Nigamem podaje, że reprezentacja unigramowa binarna jest bardziej efektywna w przypadku bardzo małych słowników, zaś reprezentacja częstościowa daje lepsze efekty dla dużych zbiorów[118]. Zaletą reprezentacji unigramowych są ich niewielkie wymagania czasowo-kosztowe. Jednakże reprezentacja taka nie pozwala odtworzyć dokumentu oryginalnego ani nawet zbliżonego do oryginału[119]. Co więcej, pominięcie szyku słów w zdaniach powoduje, że jako podobne mogą zostać wskazane dokumenty o całkowicie odmiennej treści, ale o podobnych częstościach występowania poszczególnych wyrazów, a więc reprezentowane podobnymi wektorami[120].

---

[116] Por. Ch. D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 240-241.

[117] Por. P. Gawrysiak: dz. cyt., s. 16-18.

[118] Por. A. McCallum, K. Nigam: *A comparison of event models for naive Bayes text classification*. W: *AAAI 98 workshop on learning for text categorization*. 1998. Cyt. za: P. Gawrysiak: dz. cyt., s. 35.

[119] P. Gawrysiak: dz. cyt., s. 35-36.

[120] W takim rozumieniu unigram i model unigramowy są znaczeniowo i funkcjonalnie pokrewne unitermom, rozumianym jako elementarne jednostki leksyklane jiw, pomiędzy którymi nie ma ustalonych relacji semantycznych oraz jako językom o tak zdefiniowanym słownictwie. Por. hasło *uniterm*. W: *Słownik encyklopedyczny informacji...*, s. 286.

Dla podkreślenia założenia niezależności każdego wystąpienia danego słowa model unigramowy opisywany jest wzorem na prawdopodobieństwo bezwarunkowe:

$$P_{uni}(t_1t_2t_3t_4) = P(t_1)P(t_2)P(t_3)P(t_4)$$

gdzie:

$\mathbf{t_i}$ – kolejny wyraz (ang. *term*) występujący w badanym zbiorze,

$\mathbf{P(t_i)}$ – prawdopodobieństwo wystąpienia danego wyrazu[121].

Alternatywnie, reprezentacje zachowujące kolejność wyrazów w tekście nazywane są ngramowymi. Reprezentacje takie tworzą macierze dwu- lub wielowymiarowe. Jednakże są one trudniejsze do zaimplementowania ze względu na ich wyższe wymagania technologiczne. Dla zbioru *n* słów unigram ma rozmiar równy wielkości zbioru, natomiast bigram (reprezentacja odnotowujące dwa kolejne wyrazy) może mieć rozmiar nawet n^2 słów, zaś trigram – do n^3 słów[122].

Reprezentacja dokumentów przyjęta w otwartych systemach informacyjno-wyszukiwawczych powinna umożliwiać swobodne przeszukiwanie tekstu (postulat realizowany praktycznie m.in. w mechanizmach wyszukiwarek internetowych). W związku z wymogami funkcjonalnymi takich systemów dokumenty przechowywane są w postaci jak najmniej angażującej zasoby, a jednocześnie zapewniającej możliwość swobodnego przeszukiwania treści. Jednym ze stosunkowo często wykorzystywanych sposobów przechowywania treści dokumentu jest zachowanie słów wyłonionych po przeprowadzeniu operacji optymalizacji lingwistycznej tekstu. Proces ten, w skrócie, składa się z następujących etapów:

1. Usunięcie z treści wyrazów z **listy słów mało znaczących**.
2. Sprowadzenie pozostałych wyrazów do **postaci kanonicznej** lub innej postaci ustalonej przez badacza[123].

Jednym z najprostszych, a zarazem często wykorzystywanym w nurcie IR, sposobów przechowywania dokumentów jest zachowanie zoptymalizowanej treści dokumentu bez jakichkolwiek informacji o formatowaniu. Ponieważ w tak zapisanym tekście możliwe jest wielokrotne powtórzenie

[121] Wzór na prawdopodobieństwo bezwarunkowe wystąpienia słowa w tekście w modelu unigramowym; por. Ch. D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 238.

[122] Por. P. Gawrysiak: dz. cyt., s. 36-40.

[123] Poszczególne etapy procesu optymalizacji treści dokumentu zostały opisane w dalszej części niniejszego rozdziału.

tego samego wyrazu (na różnych pozycjach), sposób ten nazywany jest zbiorem słów (ang. *bag-of-words*, BOW). Poszczególne słowa występujące w dokumencie traktowane są w tej metodzie jako niezależne od siebie, co jest zgodne z założeniami modelu unigramowego.

2.3.1. Zbiór słów (*bag-of-words*)

W modelach unigramowych pomijana jest kolejność, w jakiej słowa występują w dokumencie oraz związki zachodzące pomiędzy poszczególnymi wyrazami – przechowuje się wyłącznie informację o wystąpieniu danego słowa. Angielski termin *bag* oznacza zbiór, którego elementy mogą występować wielokrotnie. Dopełnienie terminu wskazuje na typ elementów przechowywanych w takiej formie, w tym przypadku są to słowa[124]. W języku polskim definicja powyższa opisuje termin **wielozbiór**[125].

Niewątpliwą zaletą wielozbioru jest ograniczenie rozmiaru przechowywanej w ten sposób treści w stosunku do dokumentu oryginalnego, gdzie przechowywane są dodatkowo informacje o formatowaniu.

Pominięcie kontekstu, a odnotowanie jedynie liczby wystąpień danego słowa skutkuje m.in. uznaniem za podobne wyrażeń o przeciwnych znaczeniach, a używających tych samych wyrazów. Np. zdania *Adam ma samochód* oraz *Adam nie ma samochodu* po optymalizacji (usunięciu słów bardzo częstych *ma* i *nie* oraz ustaleniu formy kanonicznej dla pozostałych słowoform) zostaną uznane za tożsame, ponieważ będą reprezentowane przez wyrażenie: *Adam samochód*. Jednakże, jak podają autorzy cytowanej już pracy *An introduction to information retrieval,* intuicja podpowiada, że dokumenty o podobnej reprezentacji wielozbiorowej mają podobną treść[126]. Traktowanie dokumentów jako zbioru słów jest wygodne obliczeniowo, natomiast nie pozwala w żaden sposób na wskazanie słów ważniejszych dla treści danego dokumentu. Ponadto poszczególne słowa mogą powtarzać się w dokumentach różniących się między sobą tematycznie.

W związku z powyższym, reprezentacje dokumentów w postaci wielozbioru znajdują zastosowanie głównie w porównywaniu dokumen-

---

[124] O *bag-of-words* por. m.in.: Ch. D. Manning, H. Schütze: dz. cyt., s. 237; Ch. D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 117; P. Jackson, I. Moulinier: dz. cyt., s. 131; D. Jurafsky, J. H. Martin: dz. cyt., s. 643.

[125] Hasło *wielozbiór*. W: *Słownik encyklopedyczny informacji...*, s. 243.

[126] Ref. za: Ch. D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 117, 120-126.

tów w celu wskazania dokumentów podobnych. Zbiór słów w postaci podstawowej nie jest jednakże zbyt wygodną ani wiarygodną podstawą do wskazywania dokumentów podobnych. W celu obniżenia kosztów systemowo-czasowych procesu porównawczego często reprezentacje tego typu przekształca się do postaci usprawniającej porównania. Jednym ze sposobów jest przygotowanie **listy frekwencyjnej** wyrazów. Jest to zestawienie słów występujących w pojedynczym dokumencie lub ich kolekcji, w którym kryterium sortowania elementów jest częstość wystąpienia.

2.3.2. Lista frekwencyjna

Frekwencyjna lista słów jest najprostszym sposobem reprezentacji treści indeksowanych dokumentów. Listę taką można uzyskać w wyniku przeprowadzenia wstępnych operacji przetwarzania tekstu, zmierzających do optymalizacji materiału tekstowego. Dla pojedynczego dokumentu zestawienie frekwencyjne może przybrać postać posortowanego spisu wszystkich haseł użytych w dokumencie, wyłonionych w wyniku operacji optymalizacji tekstu. Powtarzające się wystąpienia danej postaci kanonicznej wyrazu są sumowane, w wyniku czego poszczególne elementy listy pojawiają się tylko jeden raz, natomiast ich pozycja na liście zależy od sumy wystąpień (z zastrzeżeniem, że w obrębie podzbioru haseł o równych częstościach stosuje się dodatkowe, najczęściej alfabetyczne, kryterium sortowania). Dzięki zsumowaniu wystąpień poszczególnych elementów listy zapis fizyczny tej reprezentacji dokumentu zajmuje jeszcze mniej miejsca niż zapis podstawowej postaci wielozbiorowej[127].

Z powodu przyjętych częstościowych kryteriów sortowania listy takie nie odzwierciedlają lokalizacji (miejsca) wyrazu w tekście dokumentu (wielozbiór w podstawowej postaci rejestruje poszczególne wyrazy w kolejności ich występowania w dokumencie). Wskazanie słów kluczowych na podstawie ich frekwencyjnego zestawienia opiera się na założeniu, że leksemy o odpowiednio wysokich częstościach występowania niosą

[127] Całkowity rozmiar fizyczny zbioru przechowującego reprezentację treści dokumentu jest istotny ze względu na możliwość wczytania całej jego zawartości do pamięci operacyjnej komputera, dzięki czemu czas dostępu do danych zostaje wydajnie skrócony. W przypadku plików zbyt dużych lub zbyt dużej ich liczby, zachodzi konieczność sekwencyjnego wczytywania do pamięci operacyjnej kolejnych fragmentów pliku/plików, co znacząco wydłuża proces przetwarzania lingwistycznego treści dokumentu.

najwięcej informacji. Oczywiście z wyłączeniem z porównania słów występujących najczęściej oraz najrzadziej.

Lista frekwencyjna rejestrująca dodatkowo dane liczbowe dotyczące frekwencji poszczególnych haseł stanowi najprostszą postać tzw. **wektorowej reprezentacji dokumentu** (inaczej **wektor dokumentu**). Zaprezentowanie dokumentu, a szczególnie całej kolekcji dokumentów, w postaci ich reprezentacji wektorowych usprawnia wykonanie operacji porównawczych na dokumentach, co jest przydatne w procesie wyszukiwania dokumentów o podobnej treści. Pojedynczy wektor nie wnosi więcej informacji na temat dokumentu niż lista, ale może zostać wykorzystany do ustalenia podobieństwa pomiędzy różnymi dokumentami, co z kolei ułatwia wyszukanie dokumentów relewantnych do zapytania użytkownika.

2.3.3. Reprezentacja wektorowa

Model przestrzeni wektorowej został przedstawiony przez Gerarda Saltona jako sposób reprezentacji treści dokumentu oraz zapytania, a także metoda pozyskiwania informacji[128]. W modelu tym dokument oraz zapytanie reprezentowane są przez wielowymiarowe wektory. Każdy dokument (tekst bądź inny obiekt) jest w tej metodzie reprezentowany przez nwymiarowy wektor cech (współrzędne wektora):

$$\vec{D} = [d_1, d_2, \ldots, d_n]$$

gdzie $\mathbf{d}_i$ jest liczbą rzeczywistą.

Poszczególne wymiary *i* odpowiadają kolejnym słowom, a wartości d_i określają wagę danej cechy. Wagi poszczególnych słów określa się najczęściej jako funkcję częstości ich wystąpień w dokumencie. Każdy wymiar wektora odzwierciedla liczbę wystąpień danego wyrazu. Zbiór dokumentów można w metodzie VSM (ang. *Vector Space Model*) zaprezentować jako macierz wielowymiarową, której kolumnami są wektory reprezentujące poszczególne dokumenty wchodzące w skład kolekcji, natomiast wiersze odpowiadają kolejnym postaciom kanonicznym słów występujących w dokumentach. Na przecięciu wiersza (hasła) z kolumną (wektorem dokumentu) zawarta jest informacja o liczbie wystąpień danego hasła w konkretnym dokumencie. Najprostszą postać macierzy dokumentów stanowi

[128] Por. Ch. D. Manning, H. Schütze: dz. cyt., s. 303.

zapis określany jako *Boolean model*. Brak wystąpienia oznaczany jest cyfrą 0 (zero), natomiast wystąpienie słowa symbolizuje cyfra 1 (jeden)[129].

Z powyższego opisu wynika, że binarna reprezentacja dokumentu oprócz informacji o wystąpieniu terminów znajdujących się rzeczywiście w treści dokumentu, przechowuje również informacje o terminach występujących w innych dokumentach danego zbioru. W związku z taką konstrukcją nie tworzy się wektorów binarnych dla pojedynczych dokumentów, ale dla całej analizowanej kolekcji. W wyniku utworzenia reprezentacji binarnych uzyskujemy macierz wektorów nad zbiorem dokumentów.

Na podstawie porównania poszczególnych kolumn, za pomocą odpowiednich operacji matematycznych na wektorach liczbowych, można wskazać stopień podobieństwa pomiędzy poszczególnymi dokumentami. Model wektorowy zakłada również przekształcenie zapytania do postaci wektorowej, dzięki czemu można za pomocą porównania wektorów wskazać dokumenty o reprezentacji najbardziej zbliżonej do reprezentacji zapytania. Model ten pozwala w naturalny sposób posortować listę wskazanych dokumentów w zależności od stopnia podobieństwa do zapytania (a dokładniej: do jego wektorowej reprezentacji).

W modelu wektorowym korpus tekstów prezentowany jest często w postaci macierzy dokumenty – słowa (ang. *termdocument matrix*). Dla każdego korpusu o wielkości *d* dokumentów, dla których znana jest liczba słów *s* można utworzyć macierz $M_{d \times s}$, w której wiersze odpowiadają dokumentom, a kolumny poszczególnym słowom. Teksty poszczególnych dokumentów przechowywane są w postaci **zbioru słów**, zatem wystąpienie każdego słowa traktowane jest jako zdarzenie niezależne. Taka postać korpusu nie pozwala na zachowanie informacji o kolejności słów oraz o strukturze dokumentów. Nie ma zatem możliwości zapisania informacji o tytule, nagłówkach czy słowach kluczowych. Pomijane są również informacje o segmentacji tekstu, jak np. podział na zdania czy akapity[130]. Kolejnym ograniczeniem modelu wektorowego jest konieczność ustalenia liczby wymiarów (czyli uwzględnianych słów) macierzy M przed przystąpieniem do jej wypełniania danymi z dokumentów. Każdy nieuwzględniony na początku wyraz powoduje konieczność jego pominięcia lub przebudowania całej macierzy.

---

[129] O modelu wektorowym por. m.in. A. Mykowiecka: *Inżynieria lingwistyczna...*, s. 264, Ch. D. Manning, H. Schütze: dz. cyt., s. 298-300.

[130] Por. Ch. D. Manning, H. Schütze: dz. cyt., s. 543 i nast.

Model ten cechuje się jednak pewnymi zaletami, które powodują, że jest często wykorzystywany w aplikacjach NLP. Operacje na wektorach są łatwiejsze do przeprowadzenia za pomocą komputerów niż w przypadku innych sposobów reprezentacji dokumentów. Są one również tańsze pod względem mocy obliczeniowych i stopnia skomplikowania algorytmów niezbędnych do ich przeprowadzenia. Do typowych operacji na wektorach należą m.in. obliczanie odległości pomiędzy dwoma wektorami czy też tzw. ważenie wektorów.

Macierz dokumentów w modelu wektorowym zazwyczaj oznacza się symbolem X, a wyrażenie x_{ij} wskazuje liczę wystąpień i-tego leksemu w j-tym dokumencie. Przedstawienie charakterystyk frekwencyjnych dokumentów w postaci macierzy pozwala na dalsze, matematyczne porównanie tych dokumentów w celu wykazania podobieństw. Stwierdzamy, że dwa słowa są podobne pod względem znaczenia bądź wagi, jeżeli odpowiadające im wiersze macierzy są do siebie podobne. Oznacza to, że dane słowa występują z podobną częstością w poszczególnych dokumentach zbioru. Porównanie takie, przeprowadzone na odpowiednio licznej populacji dokumentów, pozwala wskazać potencjalne słowa kluczowe, opisujące z dużą zgodnością treść dokumentu. Dodatkowo zaś, wskazując kolumny, w których wartości dla poszczególnych, wybranych wierszy są podobne, można uzyskać listę dokumentów podobnych treściowo. Dokumenty takie mogą stanowić odpowiedź wyszukiwarki na zapytanie użytkownika, przedstawione w postaci listy słów kluczowych. Wtedy w macierzy X system wyszukuje wiersze odpowiadające poszczególnym słowom kluczowym i z nich odczytuje adresy dokumentów, których charakterystyki frekwencyjne odpowiadają kwerendzie[131].

Z kolei analizując poszczególne kolumny macierzy X na podstawie podobieństwa w zapisie odpowiednich kolumn, można wnioskować o podobieństwie dokumentów. Oprócz wspomnianego już typowania dokumentów w odpowiedzi na zapytanie użytkownika, można na tej podstawie dokonać wstępnego grupowania dokumentów (ang. *clustering*). Operując grupami dokumentów można dostarczyć użytkownikowi w odpowiedzi na jego zapytanie gotowy już zestaw dokumentów, co oczywiście skraca czas oczekiwania na odpowiedź.

[131] Oczywiście są to wewnętrzne oznaczenia wyszukiwarki pozwalające na identyfikację konkretnego dokumentu, właściwe adresy przechowywane są w osobnych strukturach danych. Dzięki temu macierz X zawiera jedynie informacje dotyczące treści dokumentów, zatem jej analiza nie jest opóźniana przez konieczność obsługi niepotrzebnych w tym przypadku danych.

Odległość dokumentów (a w zasadzie reprezentujących je wektorów) oblicza się według wzoru:

$$d(x,y) = 1 - \frac{xy}{(|x||y|)} = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Następną korzyścią z reprezentacji treści dokumentów w postaci macierzy wektorów jest możliwość dalszego ograniczenia populacji leksemów bądź słów do analizy. Można to uzyskać usuwając wszystkie wiersze, które cechują się tylko jedną pozycją niezerową. Taka sytuacja wskazuje wyrazy występujące tylko w jednym dokumencie, a więc potencjalnie nieistotne.

Reprezentacje frekwencyjne można podzielić na kilka rodzajów. Najprostsza z nich, rejestrująca sam fakt wystąpienia słowa w danym dokumencie, nazywana **binarną,** pozwala utworzyć macierz dwuwartościową, gdzie „1" oznacza wystąpienie słowa, a „0" jego brak. Taka postać macierzy nie uwzględnia wagi słów, wynikającej z częstości ich występowania. W systemach zautomatyzowanych reprezentacja binarna sprawdza się dobrze w początkowym etapie zbierania danych statystycznych w celu skonstruowania bardziej zaawansowanych reguł oceny dokumentów. Poza tym, uwzględniając liczbę słów używanych w dokumentach, pozwala uprościć wyszukanie dokumentów podobnych do siebie. Niemniej jednak, całkowite pomijanie częstości wystąpienia poszczególnych słów w dokumencie nie pozwala na próby uwzględnienia treści dokumentu w procesie indeksowania oraz wyszukiwania informacji.

Kolejny sposób – reprezentacja **logarytmiczna** w niewielkim zakresie uwzględnia częstość wystąpienia danego słowa, koncentrując się jednak nadal głównie na stwierdzeniu samego faktu wystąpienia. W reprezentacji tej każda niezerowa wartość macierzy (oznaczająca liczbę wystąpień danego słowa) zastępowana jest przez wynik następującego wyrażenia:

$$1 + \log_2 (x_{ij})$$

gdzie $\mathbf{x}_{ij}$ oznacza liczbę wystąpień danego słowa.

Reprezentacja ta, przy niewielkich nakładach obliczeniowych, pozwala uwzględnić w pewnym stopniu wagę wyrazu wynikającą z częstości jego powtórzeń w dokumencie. Jej ograniczeniem jest uwzględnianie jedynie częstości lokalnej, charakterystycznej dla danego dokumentu. Z tego też powodu metoda ta nie generuje wiarygodnych wskaźników wagi słowa,

ponieważ całkowicie pomija zachowanie danego słowa w pozostałych dokumentach z kolekcji.

Najbardziej uniwersalną metodą reprezentacji frekwencyjnej treści dokumentu jest reprezentacja **ważona logarytmiczna.** Ten sposób reprezentacji treści dokumentów opiera się na obserwacji, że wyrazy pozwalające grupować dokumenty występują w stosunkowo niewielkiej części tekstów. W związku z czym należy uwzględnić zarówno liczbę wystąpień słowa w jednym dokumencie, jak i sumaryczną frekwencję danego słowa we wszystkich tekstach w danej kolekcji. W tym rodzaju reprezentacji treści każde niezerowe wartości macierzy zastępuje się wartością wyrażenia:

$$(1 + \log_2(x_{ij})) \times \log_2(N/df_i)$$

gdzie:

N – liczba wszystkich dokumentów w kolekcji,

$\mathbf{df}_i$ - liczba dokumentów, w których występuje i-ty wyraz.

Metoda ta pozwala wskazać potencjalne słowa kluczowe dla zbioru dokumentów oraz wyznaczyć potencjalne grupy (klastry) dokumentów o podobnej treści. Podejście uwzględniające wagę słowa nie tylko w dokumencie, ale również w kolekcji podobnych dokumentów wydaje się najbardziej sensowne w przypadku automatycznego wyszukiwania informacji w dużych zbiorach dokumentów.

2.4. Wybrane sposoby określania wagi słów

Podczas tworzenia wektorowej reprezentacji dokumentu poszczególnym słowom przypisuje się ich wagę w danym dokumencie. Najprostszym sposobem ustalania wagi słów jest obliczenie częstości wystąpień w opisywanym dokumencie. W literaturze metoda ta określana jest jako *term frequency* i oznaczana $tf_{t,d}$[132]. Jednakże jest to rozwiązanie zbyt uproszczone, ponieważ zakłada, że każdy wyraz jest jednakowo ważny. Założenie powyższe nie uwzględnia naturalnie niskiej wagi słów gramatycznych (spójniki, przedrostki itp.), umieszczanych zazwyczaj na listach słów mało znaczących pod względem semantyki. Ponadto dla każdej dziedziny zbiór słownictwa charakterystycznego jest inny, co również nie zostało uwzględnione w schemacie częstości słów.

[132] Por. Ch. D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 543 i nast.

Rozwiązaniem wprowadzającym urealnienie znaczenia słowa jest porównanie częstości wystąpienia słowa w danym dokumencie do całkowitej częstości wystąpienia tego słowa we wszystkich dokumentach zbioru. Miara tej częstości oznaczana jest jako Tf_t. Jeszcze dokładniejsze urealnienie wagi danego słowa można uzyskać określając liczbę dokumentów, w których ono występuje (*document frequency* df_t) wraz ze wzrostem liczby dokumentów, w których słowo występuje, maleje waga danego słowa. W związku z powyższym często stosuje się wielkość odwrotną – *inversed document frequency* definiowaną jako:

$$idf_t = log_2 \frac{N}{df_t}$$

gdzie **N** – liczba dokumentów w zbiorze (korpusie).

Wartości idf_t są wysokie dla słów występujących rzadko, zaś niskie dla słów częstych.

Kombinacja powyższych sposobów ważenia dokumentów pozwala na wyprowadzenie wzoru tf/idf (*term frequency/inverse document frequency*) oznaczanego:

$$tf/idf_{t,d} = tf_{t,d} \, x \, idf_t = tf_{t,d \, x} \, log2 \, \frac{N}{df_t}$$

Wzór ten można przedstawić w następujący sposób:

$$waga = tf(w,d) \, x \, log_2 \frac{|D|}{df(w)}$$

gdzie:

tf(w,d) – liczba wystąpień słowa **w** w dokumencie **d**,

$tf = \frac{n_i}{\sum_k n_k}$ – frekwencja całkowita,

|D| - liczba dokumentów w zbiorze,

df(w) – liczba dokumentów, w których wystąpiło słowo *w*,

$\mathbf{n_i}$ – liczba wystąpień i-tego słowa w zbiorze,

$\sum_k n_k$ – liczba wszystkich słów w danym zbiorze.

Dla powyższego wzoru przyjęta jest następująca interpretacja:

– słowa występujące często w niewielkiej liczbie dokumentów generują maksymalne wartości wag – słowa te w zadowalającym stopniu charakteryzują treść,

- słowa występujące rzadko w małej liczbie dokumentów oraz te występujące w wielu dokumentach generują niskie wartości wag – nie pozwalają na wyraźne rozróżnienie dokumentów,
- słowa występujące w większości dokumentów generują minimalne wartości wagowe.

2.5. Optymalizacja lingwistyczna treści dokumentu

Pierwszym etapem przygotowania lingwistycznego analizowanych tekstów jest ograniczenie sumarycznej liczby wyrazów. Do osiągnięcia tego celu wykorzystuje się dwie techniki. Jedną z nich jest usunięcie z tekstu wyrazów o małej wartości informacyjnej, np. z tzw. list słów nieznaczących (ang. *stop list, stop words*). Najczęściej są to wyrazy o najwyższych oraz najniższych częstościach występowania w dokumentach danego języka, stanowiące ok. 30% słownictwa. Kolejnym sposobem jest sprowadzenie pozostałych wyrazów do jednolitej postaci, zwanej inaczej postacią kanoniczą wyrazu. W tym celu można zastosować jedną z dwóch metod – sprowadzenie wyrazów do wspólnego rdzenia, czyli tzw. *stemming* lub sprowadzenie do podstawowej formy gramatycznej (lematu, ang. *lemma*, stąd nazwa procesu: lematyzacja). Dzięki temu można ograniczyć zbiór wejściowy do indeksowania treści.

2.5.1. Przygotowanie dokumentów do indeksowania treści

Dokumenty w systemach informacyjno-wyszukiwawczych są przechowywane w postaci oryginalnej (z zachowaniem formatowania), w celu dostarczenia ich użytkownikom w odpowiedzi na zapytania. Natomiast na potrzeby samego procesu wyszukiwania dokumentów w systemie ich treść jest przetwarzana do postaci przyjętej w danym systemie i dostosowanej do procesu porównywania dokumentów oraz zapytań, stosowanego do tej kolekcji dokumentów. Pierwszy etap procesu optymalizacji polega na technicznym przygotowaniu treści archiwizowanego dokumentu do dalszej analizy. Jak wspomniano wcześniej, w trakcie kopiowania danych do lokalnych repozytoriów można automatycznie zmniejszyć rozmiar pliku oraz liczbę informacji tekstowych do opracowania na późniejszym etapie, bez żadnego wpływu na zawartą treść.

2.5.2. Usunięcie wyrazów mało znaczących

Najprostszym do przeprowadzenia, z technicznego punktu widzenia, sposobem optymalizacji rozmiaru dokumentu jest usunięcie wszystkich słów o odpowiednio wysokiej częstości występowania w dokumencie. Polega to na posortowaniu wszystkich wyrazów według liczby ich wystąpień w tekście. Zaletą takiego podejścia jest bardzo krótki czas przygotowania dokumentu do analizy treści oraz niskie koszty operacyjne i czasowe. Operację sortowania można przeprowadzić na zasadzie procesu segmentacyjnego, w momencie wydzielania poszczególnych tokenów z treści, trywialnego dla systemu operacyjnego. W celu uzyskania większej wiarygodności oraz pewności, że usunięte są jedynie wyrazy nieistotne należy już na tym etapie, przed określeniem częstości, sprowadzić wyrazy do jednej formy gramatycznej. Jak podaje np. A. Mykowiecka, słowa o bardzo wysokich frekwencjach pojawiają się w znacznej większości dokumentów danego języka, w związku z czym nie są odpowiednimi kandydatami do reprezentowania treści jako słowa kluczowe. Z kolei słowa o najniższych częstościach występowania pojawiają się w nielicznych dokumentach, w dodatku są to wystąpienia sporadyczne, niecharakteryzujące w żaden sposób treści dokumentów[133].

Kolejnym sposobem jest usunięcie z tekstu wyrazów z tzw. strefy słownictwa gramatycznego. Dla przypomnienia: należą do niej hasła o największych częstościach, są to: podstawowe proste przyimki (takie jak: w, z, za), spójniki (że, i), zaimki (on) oraz czasowniki posiłkowe (być). Rozwiązanie to oferuje większą wiarygodność niż poprzednie, ponieważ listy słownictwa gramatycznego tworzone są na podstawie analizy dużych korpusów tekstów. Materiał do zbudowania takich list musi być reprezentatywny dla danego języka albo przynajmniej dla danej dyscypliny, co więcej, przed dodaniem na listę dane słowo musi zostać zaakceptowane przez człowieka, jako spełniające kryteria przynależności do danej strefy leksykalnej. Oczywiście, pozycje na takich listach przedstawiane są w znormalizowany sposób. Zatem treść przetwarzanych dokumentów również powinna zostać poddana operacji sprowadzenia do jednolitej postaci. Metoda ta nie jest powszechna wśród wyszukiwarek internetowych, ponieważ najlepsze rezultaty pod względem wartości danego słowa uzyskuje się tworząc listy słownictwa gramatycznego dla poszczególnych dys-

[133] Por. A. Mykowiecka: *Inżynieria lingwistyczna...*, s. 261.

cyplin. Oczywiście odpowiednie przygotowanie takich zestawień jest procesem zarówno czaso-, jak i zasobochłonnym, czyli mniej opłacalnym dla mechanizmów wyszukiwawczych.

Ostatnim sposobem, częściej stosowanym, jest skorzystanie z list słów nieznaczących. Są to słowniki zawierające zestawienia wyrazów występujących najczęściej w tekstach danego języka. Zazwyczaj tworzy się je w oparciu o statystyki frekwencyjne języka[134]. Przykładowe stop listy można znaleźć pod następującymi adresami:

- lista dla języka polskiego, serwis Ranks.nl[135]: http://www.ranks.nl/stopwords/polish.html,
- lista dla języka polskiego, serwis Wikipedia[136]: http://pl.wikipedia.org/wiki/Wikipedia:Stopwords,
- listy dla języków europejskich, dostępne w serwisie Ranks.nl[137]: http://www.ranks.nl/resources/stopwords.html,
- lista dla języka angielskiego przygotowana przez firmę MySql[138]: http://dev.mysql.com/doc/refman/5.0/en/fulltext-stopwords.html.

Zestawienia takie jest stosunkowo łatwo przygotować, ponieważ nie uwzględnia się w nich funkcji danego słowa, a jedynie na częstość jego wystąpienia w tekstach języka. Dzięki temu listy słów nieznaczących mogą zostać przygotowane automatycznie. Zależnie od metodologii, listy słów mało znaczących mogą być przygotowywane dla odpowiednich wystąpień danego słowa lub, po standaryzacji, dla wybranej formy wzorcowej. Jednakże, ponieważ proces eliminacji z tekstu wyrazów nieznaczących jest jednym z pierwszych etapów analizy lingwistycznej tekstu przeprowadzanym na oryginalnej wersji tekstu, listy słów mało znaczących rejestrują konkretną postać fleksyjną danego wyrazu. Lista wyrazów mało znaczą-

---

[134] Listy te jednak nie są standardami, ale propozycjami do wykorzystania. Poza tym należy pamiętać, że język naturalny jest systemem ciągle się rozwijającym, zatem zestawienia słów nieznaczących powinny być na bieżąco aktualizowane.

[135] Por. *Polish stopwords* [on-line]. [Dostęp: 15 paździeirnika 2011]. Dostępny w World Wide Web: http://www.ranks.nl/stopwords/polish.html.

[136] Por. *Wikipedia:stopwords* [on-line]. [Dostęp: 15 października 2011]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/Wikipedia:Stopwords.

[137] Por. *English stopwords* [on-line]. [Dostęp: 15 października 2011]. Dostępny w World Wide Web: http://www.ranks.nl/resources/stopwords.html. Zestawienia dla poszczególnych języków dostępne są po wybraniu odpowiedniego odnośnika znajdującego się u dołu strony (np. French dla języka francuskiego).

[138] Por. *11.8.4. Full-Text Stopwords*. W: *MySQL 5.0 Reference Manual* [on-line]. [Dostęp: 15 października 2011]. Dostępny w World Wide Web: http://dev.mysql.com/doc/refman/5.0/en/fulltext-stopwords.html.

cych zastosowana w operacjach przygotowania tekstów na potrzeby niniejszej książki została zaprezentowana w tabeli 4.

Przefiltrowanie treści dokumentu pozwala zredukować rozmiar indeksu haseł występujących w dokumencie do około 50%. Należy jednak mieć świadomość, że wyrazy usunięte w wyniku filtrowania słów nieznaczących nie będą w dokumencie wyszukiwane, więc niektóre zapytania użytkowników mogą pozostać bez odpowiedzi[139].

Nadmierne zredukowanie liczby wyrazów w treści dokumentu ma również wpływ na swobodne wyszukiwanie, a także na wyszukiwanie fraz lub zdań, ponieważ brak wyrazów funkcyjnych zmienia treść zdania. W związku z tym można zaobserwować tendencję do zmniejszania rozmiarów list słów nieznaczących w systemach informacyjno-wyszukiwawczych z kilkuset potencjalnych słów do kilku lub kilkunastu wyrazów. W przypadku filtrowania z treści słów nieznaczących można również uwzględnić frekwencje współwystępowania poszczególnych wyrazów i pozostawić te częste wyrazy, które wielokrotnie występują w połączeniu z innymi, bardziej znaczącymi wyrazami[140].

Po odfiltrowaniu z tekstu wyrazów nieznaczących można przejść do kolejnego etapu, którym jest ujednolicanie postaci uzyskanych słów. Jak już wcześniej wspomniano, można w tym celu posłużyć się procedurą wyznaczania rdzenia wyrazu lub sprowadzania do podstawowej formy gramatycznej (postać kanoniczna, hasło) w procedurze lematyzacji.

2.5.3. Wyznaczanie rdzenia wyrazu

Można wskazać dwie metody sprowadzania wyrazów do jednolitej postaci. Metodą prostszą, łatwiejszą do zaimplementowania jest stemming. W wyniku operacji sprowadzania do wspólnego rdzenia powstaje tzw. rdzeń wyrazu (ang. *stem*), czyli ciąg liter niezmiennych dla podobnych graficznie wyrazów[141]. Najczęściej postać taką uzyskuje się w wyniku usunięcia z wyrazów ich sufiksów oraz końcówek fleksyjnych (określających

---

[139] Por. Ch. D. Manning, H. Schütze: dz. cyt., s. 533-534, Ch. D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 2728, 233, P. Jackson, I. Moulinier: dz. cyt., s. 33, 66.

[140] Por. Ch. D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 27. A. Mykowiecka: *Inżynieria lingwistyczna...*, s. 261.

[141] Por. A. Mykowiecka: tamże, s. 69, 80; Ch. D. Manning, H. Schütze: dz. cyt., s. 132-134, 534; D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 32-34; D. Jurafsky, J. H. Martin: dz. cyt., s. 58.

Tabela 4. Rozbudowana lista słów nieznaczących dla języka polskiego stosowana na potrzeby niniejszej książki.

co	ich	jestem	mną	nie	powinni	tej	w
cokolwiek	ile	jeszcze	mnie	niego	powinno	ten	w
coś	im	jeśli	mogą	niej	poza	teraz	w
czasami	inna	jeżeli	moi	niemu	prawie	też	w
czasem	inne	już	moim	nigdy	przecież	to	w
czemu	inny	ją	moja	nim	przed	tobą	w
czy	innych	każdy	moje	nimi	przede	tobie	w
czyli	iż	kiedy	może	niż	przedtem	toteż	w
daleko	ja	kilka	możliwe	no	przez	trzeba	w
dla	ją	kimś	można	o	przy	tu	w
dlaczego	jak	kto	mój	obok	roku	tutaj	w
dlatego	jakaś	ktokolwiek	mu	od	również	twoi	w
do	jakby	ktoś	musi	około	sam	twoim	z
dobrze	jaki	która	my	on	sama	twoja	za
dokąd	jakichś	które	na	ona	są	twoje	za
dość	jakie	którego	nad	one	się	twym	za
dużo	jakiś	której	nam	oni	skąd	twój	ze
dwa	jakiż	który	nami	ono	sobie	ty	zn
dwaj	jakkolwiek	których	nas	oraz	sobą	tych	zn
dwie	jako	którym	nasi	owszem	sposób	tylko	zo
dwoje	jakoś	którzy	nasz	pan	swoje	tym	ża
dziś	je	ku	nasza	pana	są	u	ża
dzisiaj	jeden	lat	nasze	pani	ta	w	ża
gdy	jedna	lecz	naszego	po	tak	wam	ża
gdyby	jedno	lub	naszych	pod	taka	wami	że
gdyż	jednak	ma	natomiast	podczas	taki	was	że
gdzie	jednakże	mają	natychmiast	pomimo	takie	wasz	
gdziekolwiek	jego	mam	nawet	ponad	także	wasza	
gdzieś	jej	mi	nią	ponieważ	tam	wasze	
go	jemu	mimo	nic	powinien	te	we	
i	jest	między	nich	powinna	tego	według	

owanie własne na podstawie *Polish stopwords, large list* [on-line]. [Dostęp: 15 października 2011]. Dostępny w Wo
nks.nl/stopwords/polish.html oraz wyników doświadczeń.

funkcję składniową wyrazu w wypowiedzeniu). Ze względu na stosunkową łatwość zaimplementowania procedury stemmingu w postaci algorytmu oraz prostotę samej operacji wskazywania wspólnej formy wyrazu, jest to najczęściej stosowana w wyszukiwarkach metoda optymalizacji poszczególnych dokumentów do indeksowania ich treści. Procedura wskazywania rdzenia wyrazu jest szczególnie przydatna w przypadku języków o prostej fleksji.

Istnieje wiele algorytmów wskazujących wspólny rdzeń, szczególnie dla języka angielskiego, m.in. *Lovins Stemmer, Porter stemming algorithm, Paice/Husk alogrithm*. Najpowszechniej wykorzystywanym dla przetwarzania tekstów angielskich jest obecnie algorytm Portera (dokładniej Porter2), stanowiący aktualnie część projektu SNOWBALL – języka tworzenia narzędzi do wyszukiwania rdzeni wyrazów dla wielu języków naturalnych[142]. Narzędzia dla języka polskiego zostały opisane w dalszej części niniejszej rozprawy.

2.5.3.1. Metody wskazywania wspólnego rdzenia

Dla dokumentów języka angielskiego, najpopularniejszym i jednocześnie najtańszym sposobem optymalizacji danych wejściowych jest wskazanie wspólnych rdzeni dla wyrazów. Wynika to z faktu, że wyszukanie wspólnego rdzenia jest operacją mechaniczną, wykorzystującą zapis graficzny (lub dźwiękowy) danego wyrazu. W jej trakcie nie ma potrzeby pozyskiwania jakichkolwiek informacji o właściwościach gramatycznych wyrazu czy też o jego kontekście w wyrażeniu. Dzięki temu uproszczeniu można uzyskać duże oszczędności czasu, jak i nakładu pracy podczas opracowywania tekstów.

Dwie najczęściej spotykane metody wskazania wspólnego rdzenia to metoda słownikowa oraz metoda algorytmiczna. Najogólniej – metoda słownikowa opiera się na słownikach wzorcowych rdzeni wyrazów. Jej implementacja polega na porównaniu odpowiedniej części słowa (liczonej w znakach) z zawartością słownika i na podstawie stwierdzenia podobień-

[142] Więcej o algorytmie Portera por. *The Porter Stemming Algorithm* [on-line]. [Dostęp: 15 października 2011]. Dostępny w World Wide Web: http://tartarus.org/~martin/PorterStemmer/; witryna projektu SNOWBALL [on-line]. [Dostęp: 19 października 2011]. Dostępny w World Wide Web: http://snowball.tartarus.org/; projekt Oleander Solutions [on-line]. [Dostęp: 19 października 2011]. Dostępny w World Wide Web http://www.oleandersolutions.com/stemming/stemming.html.

stwa wskazaniu poprawnego rdzenia dla danego słowa. Metoda ta w postaci typowej nie rozwiązuje np. problemu homonimii, jednakże stosując dodatkowo informacje o częstości poszczególnych rdzeni w odpowiednio bogatym korpusie można z dużym prawdopodobieństwem wskazać poprawny rdzeń. W metodzie algorytmicznej natomiast, w oparciu o analizę zadanego korpusu, tworzony jest zestaw reguł transformacji postaci wyrazu do rdzenia. Metoda ta, w odróżnieniu od słownikowej, pozwala przetwarzać również wyrazy niezawarte w słowniku wzorcowym, jednakże w przypadku języków fleksyjnych może prowadzić do tworzenia rdzeni nieprawidłowych. Dodatkową zaletą metod algorytmicznych jest możliwość uzyskania większej wydajności niż w przypadku porównywania ze słownikiem. Natomiast najlepsze efekty uzyskać można łącząc obie metody w tzw. metodę hybrydową. Kosztem pewnego spowolnienia działania programu zyskuje się znacznie większą precyzję oraz możliwość prawidłowego przetwarzania nowych wyrazów, nieznajdujących się jeszcze w słowniku[143].

Dla języka polskiego można wyróżnić dwa projekty podejmujące problem wskazania wspólnego rdzenia dla wyrazów. Metodą bazująca na algorytmach z pewnym wsparciem z własnych słowników jest projekt Andrzeja Białeckiego – *STEMPEL – Algorytmiczny Stemmer dla języka polskiego*[144]. Projekt dostarcza własne słowniki rdzeni, jednakże, według słów jego autora, w ok. 74% dane te są komercyjne.

Kolejny projekt wskazywania rdzeni dla języka polskiego jest dostępny bez opłat licencyjnych. Jest to pakiet morfologik-stemming (dawniej Stemplator/Lametyzator)[145] autorstwa Dawida Weissa, dostępny obecnie jako część projektu Morfologik[146]. Właściwie pakiet ten jest biblioteką napisaną w języku JAVA, łączącą funkcje stemmera słownikowego oraz lematyzatora. Rozbudowane możliwości analityczne w zakresie składni

[143] O polskich aplikacjach wskazujących wspólny rdzeń wyrazów (stemmerach) por. D. Weiss: *A survey of freely available polish stemmers and evaluation of their applicability in information retrieval* [on-line]. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://www.cs.put.poznan.pl/dweiss/site/publications/download/ltc_092_weiss_2.pdf.

[144] Por. A. Białecki: *Stempel – algorithmic stemmer for Polish language* [on-line]. [Dostęp: 19 października 2011]. Dostępny w World Wide Web: http://getopt.org/stempel/index.html#distrib.

[145] Por. D. Weiss: *Dawid Weiss – Lematyzator dla języka polskiego* [on-line]. Poznań: Politechnika Poznańska, Zakład Inteligentnych Systemów Wspomagania Decyzji 2006. [Dostęp: 19 października 2011]. Dostępny w World Wide Web: http://www.cs.put.poznan.pl/dweiss/xml/projects/lametyzator/index.xml?lang=pl.

[146] Por. M. Miłkowski: *Morfologik* [on-line]. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://morfologik.blogspot.com/.

i morfologii oraz darmowy charakter czynią z tej biblioteki niezwykle interesujące narzędzie przetwarzania tekstów języka polskiego. Algorytm wskazywania rdzeni bazuje na publicznie dostępnym słowniku języka polskiego programu ISPELL[147] oraz bibliotece FSA (*Finite State Automata*, Automatów Skończonych) autorstwa Jana Daciuka[148].

W przypadku języków fleksyjnych, do których zdecydowanie należy język polski, mechaniczne odcinanie sufiksów nie pozwala na uzyskanie poprawnych form gramatycznych (leksemów). Złożone reguły gramatyczne rządzące fleksją języka polskiego powodują dużo trudności w utworzeniu typowego algorytmu wskazywania poprawnej formy rdzeniowej. W związku z powyższym, a szczególnie w związku z bogatą fleksją języka polskiego, konieczne jest uwzględnienie nie tylko postaci graficznej danego termu, ale również jego funkcji gramatycznej oraz roli w wypowiedzeniu. Proces wskazywania podstawowej formy wyrazu na podstawie analizy morfologicznej określany jest jako lematyzacja[149].

2.5.4. Wskazywanie lematu słowoformy

Przekształcenie wyrazu prowadzące do uzyskania kanonicznej formy leksemu (reprezentacji słownikowej danego leksemu) z dowolnej jego formy gramatycznej nazywamy lematyzacją, a uzyskaną postać – lemmą lub lematem. Dopasowanie wyrazu, jako słowoformy, do postaci kanonicznej wymaga uwzględnienia jego znaczenia. Odpowiednie znaczenie można rozpoznać na podstawie kontekstu, w jakim dana słowoforma została zastosowana, co dla systemów automatycznego opracowania języka naturalnego jest trudne[150].

Podobnie, jak w przypadku wskazywania rdzenia dla wyrazów, tak i proces lematyzacji może zostać przeprowadzony na podstawie algorytmu rozpoznającego znaczenie wyrazu i wskazującego jego formę podstawową lub na podstawie operacji słownikowych. Metoda słownikowa polega na odszukaniu zbioru słowoform (leksemu) danego hasła i wskazaniu

---

[147] Por. M. Prywata: *Oficjalna strona polskiego zbioru słów dla ispella* [on-line]. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://ispell-pl.sourceforge.net/.

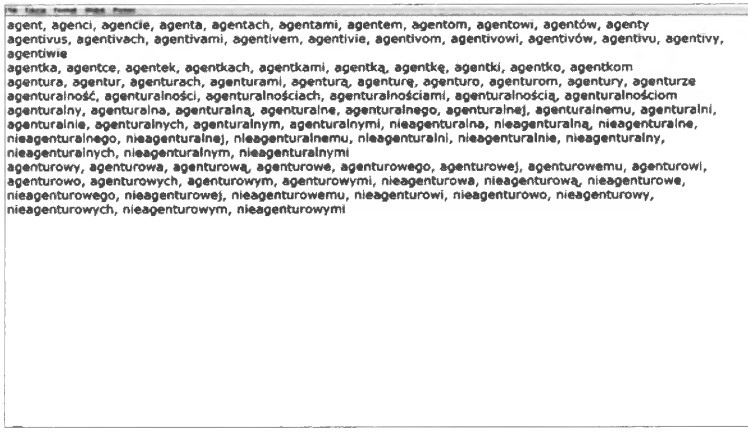
[148] Por. J. Daciuk: *Narzędzia do automatów skończonych* [on-line]. Gdańsk: Politechnika Gdańska, Katedra Inżynierii Wiedzy. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://www.eti.pg.gda.pl/katedry/kiw/pracownicy/Jan.Daciuk/personal/fsa_polski.html.

[149] Por. m.in. A. Mykowiecka: *Inżynieria lingwistyczna...*, s. 69.

[150] Por. D. Manning, P. Raghavan, H. Schütze: dz. cyt., s. 32-34.

jego postaci kanonicznej. Podejście algorytmiczne wymaga dużych mocy obliczeniowych i cechuje się wysokimi kosztami systemowymi. Jednakże nie zawsze zachodzi konieczność zastosowania skomplikowanych operacji analizy morfologicznej. W wielu przypadkach wystarczająca okazuje się metoda językowa. Dotyczy to szczególnie tych wyrazów, w przypadku których nie zachodzi problem ujednoznacznienia znaczenia. Metoda słownikowa jest łatwiejsza w implementacji, natomiast może zostać zastosowana jedynie do wyrazów umieszczonych w słowniku. Dlatego też najlepsze efekty można uzyskać łącząc obie metody i stosując analizę morfologiczną w przypadkach, gdy słowo jest niejednoznaczne lub nie ma go w słowniku. Jednym ze źródeł słownikowych, które można wykorzystać podczas lematyzacji wyrazów tekstu jest *Słownik odmian*. Fragment *Słownika* został zaprezentowany na ilustracji 1.

Ilustracja 1. Przykładowy fragment zawartości *Słownika odmian.*


agent, agenci, agencie, agenta, agentach, agentami, agentem, agentom, agentowi, agentów, agenty
agentivus, agentivach, agentivami, agentivem, agentivie, agentivom, agentivowi, agentivów, agentivu, agentivy, agentiwie
agentka, agentce, agentek, agentkach, agentkami, agentką, agentkę, agentki, agentko, agentkom
agentura, agentur, agenturach, agenturami, agenturą, agenturę, agenturo, agenturom, agentury, agenturze
agenturalność, agenturalności, agenturalnościach, agenturalnościami, agenturalnością, agenturalnościom
agenturalny, agenturalna, agenturalną, agenturalne, agenturalnego, agenturalnej, agenturalnemu, agenturalni, agenturalnie, agenturalnych, agenturalnym, agenturalnymi, nieagenturalna, nieagenturalną, nieagenturalne, nieagenturalnego, nieagenturalnej, nieagenturalnemu, nieagenturalni, nieagenturalnie, nieagenturalny, nieagenturalnych, nieagenturalnym, nieagenturalnymi
agenturowy, agenturowa, agenturową, agenturowe, agenturowego, agenturowej, agenturowemu, agenturowi, agenturowo, agenturowych, agenturowym, agenturowymi, nieagenturowa, nieagenturową, nieagenturowe, nieagenturowego, nieagenturowej, nieagenturowemu, nieagenturowi, nieagenturowo, nieagenturowy, nieagenturowych, nieagenturowym, nieagenturowymi

Źródło: Opracowanie własne na podstawie danych ze *Słownika odmian*[151].

Zoptymalizowana treść dokumentów stanowi podstawę do dalszych operacji na dokumentach. Przykłady takich operacji na tekstach przygotowanego korpusu badawczego zostały opisane w następnym rozdziale.

[151] Por. *Słownik SJP.pl – odmiany słów* [on-line]. [Dostęp: 12 października 2010]. Dostępny w World Wide Web: http://www.sjp.pl/slownik/odmiany/.

Zasady postępowania badawczego i opis przygotowanego systemu

Zastosowanie komputerów jest oczywistym wyborem dla jakichkolwiek operacji przetwarzania oraz porównywania dowolnych zestawów danych. Sposób oraz zasady przeprowadzania tych operacji ustalane są przez człowieka, natomiast przetwarzanie maszynowe pozwala na wykonanie wielu pojedynczych operacji w bardzo krótkim czasie, co przyczynia się do zaoszczędzenia czasu i kosztów oraz zapewnia większą precyzję i dokładność obliczeń. Nic więc dziwnego, że techniki komputerowe znalazły tak szerokie zastosowanie w przetwarzaniu tekstów języka naturalnego, gdzie badacze mają do czynienia z korpusami wyrażeń językowych liczącymi miliony elementów. Operacje dokonywane podczas analizy tekstów to wybór poszczególnych tokenów (wyrazów), sumowanie liczby wystąpień pojedynczych wyrazów, ewentualne sprowadzenie wszystkich słów do ustalonej postaci wzorcowej oraz porównania statystyczne przeprowadzane w obrębie danych tekstowych dla jednego dokumentu bądź całej kolekcji.

Również badania przeprowadzone w ramach niniejszej książki w szerokim zakresie wykorzystywały moce obliczeniowe oraz pamięć komputerów.

Wszelkie operacje na tekstach analizowanych dokumentów przeprowadzane były zgodnie z zasadami NLP. Bieżący rozdział prezentuje zasady postępowania badawczego, materiał wykorzystany podczas badań oraz aplikację komputerową przygotowaną w celu przeprowadzenia założonych badań.

3.1. Przedmiot, cel i metodologia badań

Ponieważ w trakcie badań nad tekstami języka naturalnego można podejmować różne, często dosyć odległe tematycznie problemy badawcze, każdy rodzaj badań wymaga odrębnych procedur postępowania. Pracę bieżącą poświęcono zagadnieniom związanym z nurtem statystycznym przetwarzania języka naturalnego, w związku z czym zastosowana metodologia badań, procedury i operacje badawcze dostosowane są do potrzeb analizy statystycznej tekstów języka naturalnego.

3.1.1. Przedmiot badań

Przedmiotem przeprowadzonych badań jest praktyczne zastosowanie odkryć i prawideł nurtu statystycznego NLP. W rozdziale 1., *Związki NLP z informacją naukową,* zaprezentowana została geneza rozwoju badań w zakresie przetwarzania języka naturalnego. Wskazano tam, że jednym z dwóch głównych kierunków badawczych NLP jest analiza statystyczna tekstów. Osiągnięcia badań statystycznych nad językiem naturalnym zostały wykorzystane m.in. w nurcie NLP znanym jako IR (ang. *Information Retrieval*) do usprawnienia wyszukiwania dokumentów w dużych zbiorach, najczęściej przez systemy informacyjno-wyszukiwawcze. Usprawnienie w tym przypadku dotyczy zarówno jakości i wartości odpowiedzi udzielanych przez system na zapytania użytkowników, jak i samego procesu przygotowania reprezentacji treści dokumentów przechowywanych w systemie.

Badania zostały przeprowadzone bezpośrednio na wybranych tekstach języka polskiego. W ich trakcie wykorzystano również dyspozycję kognitywną człowieka w zakresie tworzenia reprezentacji treści za pomocą słów kluczowych. Ze względu na cel pracy, przedmiotem badań był również autorski system automatycznej analizy tekstów języka polskiego oraz generowania słów kluczowych dla tych tekstów.

Zakres badań w niniejszej książce jest ograniczony do zagadnień związanych z usprawnieniem wyszukiwania dokumentów w systemach informacyjno-wyszukiwawczych na podstawie odpowiedniej reprezentacji treści dokumentów. Przeprowadzone i omówione tu badania, podobnie jak cały zakres badawczy przetwarzania tekstów języka naturalnego, mają charakter interdyscyplinarny. W zakresie badawczym niniejszej książki mieszczą się metody i zasady indeksowania oraz problematyka reprezen-

tacji treści dokumentów, dostarczane przez informację naukową. W trakcie prac badawczych wykorzystane zostały również rozwiązania informatyczne – w zakresie wyposażenia języków programowania w możliwości analizy tekstów. Przeprowadzono także analizę wyznaczania słów kluczowych jako charakterystyk treści dokumentów.

3.1.2. Cele i hipotezy badawcze

Głównym celem badań stało się porównanie skuteczności metod automatycznych i tradycyjnych w tworzeniu charakterystyk wyszukiwawczych dokumentów, ze szczególnym uwzględnieniem słów kluczowych. Ponadto założono przeprowadzenie badania i oceny możliwości automatycznego generowania słów kluczowych jako reprezentacji treści dokumentów.

Systemy informacyjno-wyszukiwawcze operują na reprezentacjach treści dokumentów, do których dostęp potrzebny jest użytkownikom. Jak podaje Barbara Sosińska-Kalata, dokumenty w danym zbiorze (niekoniecznie w systemie informacyjno-wyszukiwawczym) odwzorowują jakiś zakres wiedzy, zaś ich opisy, w postaci metainformacji związanej z dokumentem, stanowią symboliczne reprezentacje owych fragmentów wiedzy publicznej. Przy takich założeniach, autorka konkluduje, że zadanie systemów dokumentacyjnych polega na odwzorowaniu fragmentów wiedzy w dokumentach – właśnie za pomocą metadanych. Sprowadza się to do operacji dopasowania metainformacji do zapytań użytkowników i wskazaniu na tej podstawie odpowiednich dokumentów[152]. Słowa kluczowe odwzorowując treść dokumentu tworzą jednostkę metainformacji, charakterystykę słowną dokumentu. Charakterystyki słowne powstają m.in. w trakcie procesu opracowania rzeczowego dokumentu na potrzeby katalogowania. Podczas analizy treści dokumentu prowadzonej przez człowieka wskazuje się wyrażenia charakteryzujące treść danej pozycji[153].

Przygotowanie prawidłowej reprezentacji, wyrażonej w odpowiednim, stosowanym w danym systemie języku informacyjno-wyszukiwawczym wymaga zarówno wprawy, ze względu na konieczność opanowania leksyki i gramatyki danego języka, jak i czasu, z powodu wymogu zapoznania się z treścią dokumentu i przygotowania odpowiedniego opisu owej

[152] Za: B. Sosińska-Kalata: dz. cyt., s. 29.

[153] O funkcji słów kluczowych w opisie rzeczowym dokumentu oraz o sposobach wyznaczania takich słów por. m.in. T. Głowacka: *Analiza dokumentu i jego opis przedmiotowy*. Warszawa: SBP 2003, s. 15-16.

treści. Możliwość zautomatyzowania tego procesu niesie ze sobą niewątpliwe korzyści w postaci oszczędności kosztów pracy oraz czasu związanych z opracowaniem odpowiedniej charakterystyki. Dlatego też wszelkie badania przybliżające praktyczne wdrożenie takich możliwości mogą przynieść bardzo cenne wyniki.

Systemami informacyjno-wyszukiwawczymi operującymi na największych zbiorach dokumentów są obecnie wyszukiwarki internetowe. Indeksują one zbiory liczące miliardy egzemplarzy, co stanowi przynajmniej kilkukrotność zasobów indeksowanych w zamkniętych systemach informacyjno-wyszukiwawczych (jakimi są np. katalogi biblioteczne online). Ze względu na coraz większą popularność dostępu i wykorzystania zasobów internetowych, wśród użytkowników tej formy dostępu do informacji wykształciły się zachowania wyszukiwawcze dostosowane do mechanizmów jakimi przeszukują oni zasoby sieci Web. Najpowszechniejszym obecnie sposobem zadawania zapytań wyszukiwawczych jest przedstawienie ich w postaci listy słów kluczowych reprezentujących treść poszukiwanego dokumentu[154]. Tworząc słowa kluczowe użytkownicy nie są ograniczeni do wyboru słownictwa specjalistycznych słowników, obowiązkowych przy wyszukiwaniach w zamkniętych systemach wyszukiwawczych. Swoboda tworzenia własnych zestawów słów kluczowych na podstawie słownictwa niekontrolowanego oraz możliwość wskazania przez wyszukiwarkę dokumentów odpowiednich dla danego zapytania, dostępnych w sieci Web, są przyczynami rosnącej popularności właśnie takiego sposobu przeszukiwania zasobów.

Uwzględniając trudność w tworzeniu poprawnych formalnie zapytań wyszukiwawczych oraz stosunkową łatwość i swobodę w przypisywaniu słów kluczowych, a także rosnącą popularność tej metody oznaczania treści poszukiwanych dokumentów wśród użytkowników, w niniejszej książce największą uwagę skierowano właśnie na słowa kluczowe.

---

[154] Poszczególne wyszukiwarki publikują zestawienia najpopularniejszych słów kluczowych używanych przez internautów, np. dla wyszukiwarki Google.com lista najpopularniejszych słów kluczowych w wyszukiwaniach dla języka polskiego dostępna jest pod adresem: *Statystyki wyszukiwarki Google* [on-line]. [Dostęp: 27 grudnia 2010]. Dostępny w World Wide Web: http://www.google.com/insights/search/#geo=PL&date=1%2F2010%2012m&cmpt=q; dla wyszukiwarki serwisu Wirtualna Polska (obsługiwanej przez mechanizm NetSprint) lista szczegółowa dostępna jest pod adresem: *Najpopularniejsze zapytania* [on-line]. [Dostęp: 27 grudnia 2010]. Dostępny w World Wide Web: http://szukaj.wp.pl/najpop.html.

W związku z postawionymi celami podjęto następujące konkretne problemy badawcze:

1. Czy można zautomatyzować proces tworzenia charakterystyk wyszukiwawczych dokumentów z zawężeniem do słów kluczowych?
2. Czy słowa kluczowe wskazane w wyniku analizy frekwencyjnej treści dokumentu będą zgodne ze słowami ustalonymi przez człowieka?

Kierując się dotychczasowym doświadczeniem w zakresie badań lingwistyki kwantytatywnej autor postawił kilka hipotez roboczych, które zostały zweryfikowane w trakcie przeprowadzonych badań. Rozważone zostały następujące hipotezy:

1. Po usunięciu z treści dokumentu naukowego słów mało znaczących[155], dla pozostałych słów, sprowadzonych do postaci podstawowej oraz posortowanych ze względu na liczbę wystąpień w dokumencie, w obrębie leksemów o najwyższych frekwencjach można wskazać potencjalne słowa kluczowe.
2. Przy ustalaniu wag słów na potrzeby wskazania słów kluczowych wystarczającą miarą będzie stosunek liczby wystąpień danego leksemu do całkowitej częstości tego leksemu w całym zbiorze dokumentów. Dzięki potwierdzeniu tego założenia, można by znacznie obniżyć koszty operacyjne automatycznego wskazywania słów kluczowych, ponieważ najpopularniejszą chyba metodą wskazywania wagi słowa stosowaną współcześnie jest obliczanie współczynnika zwanego tf-idf, co wymaga przeprowadzenia bardziej skomplikowanych obliczeń oraz operacji na etapie przygotowywania dokumentów do analizy.
3. Wyróżnienie przez autora danego słowa w treści dokumentu pozwala na zwiększenie wagi odpowiedniego leksemu na liście frekwencyjnej leksemów.
4. Ze względu na kontekstowość przygotowania przez autorów słów kluczowych oraz uwzględnianie w wyrażeniach tego typu również elementów pozatreściowych, programy komputerowe nie są w stanie zaproponować równie dobrze dopasowanych haseł.

---

[155] Zdefiniowanych zgodnie z zasadami podanymi w r. *Sposoby określania wagi poszczególnych słów* oraz w r. *Usunięcie wyrazów mało znaczących*. Zestawienie zastosowanych w bieżącej pracy słów mało znaczących zawiera tabela 4. *Rozbudowana lista słów nieznaczących dla języka polskiego.*

W celu realizacji przyświecających niniejszej książce założeń i związanych z nimi hipotezami badawczymi przeprowadzono szereg badań i analiz, które zostały opisane w podrozdziale poświęconym organizacji i przebiegowi badań.

3.2. Zastosowane metody, technologie i narzędzia badawcze

Ze względu na oczywiste korzyści wynikające z zastosowania komputerów i programów komputerowych do przetwarzania i analizy danych, na potrzeby bieżącej pracy autor opracował i przygotował program komputerowy w postaci skryptów języka Python, wspierający przeprowadzenie zaplanowanych prac badawczych.

3.2.1. Zastosowane technologie

Każdą aplikację pracującą w systemie komputerowym tworzy się w języku programowania, za pomocą którego prezentuje się problem oraz sposób jego rozwiązania przy użyciu komputera. Języki programowania są to języki sztuczne, wytworzone intencjonalnie w konkretnym celu (w odróżnieniu od języków naturalnych, etnicznych), których słownictwo i gramatyka są ściśle i jednoznacznie zdefiniowane. Słownik języka programowania składa się z instrukcji i poleceń związanych z przetwarzaniem danych oraz obsługą komunikacji człowiek-komputer i aplikacja-komputer, natomiast gramatyka określa zasady i skutki zastosowania poszczególnych elementów dostępnego słownictwa[156].

Wybór języka programowania do stworzenia aplikacji mającej na celu wsparcie analizy tekstów języka naturalnego zależy w dużej mierze od indywidualnych preferencji i możliwości programisty. Sprawną i skuteczną aplikację można stworzyć wykorzystując większość dostępnych obecnie języków programowania. Wybierając konkretne środowisko programiści kierują się swoim poziomem znajomości danego języka (lub języków podobnych) oraz wsparciem zarówno ze strony twórców języka, jak i środo-

[156] Opracowanie własne na podstawie hasła *język programowania*. W: *Encyklopedia PWN* [on-line]. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://encyklopedia.pwn.pl/haslo.php?id=3917948, hasło *języki programowania*. W: *isłownik.pl (słownik slangu informatycznego)* [on-line]. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://www.i-slownik.pl/1,738,jezyki,programowania.html.

wiska innych programistów. Ważnym aspektem jest również liczba i dostępność tzw. **bibliotek**, czyli gotowych, działających fragmentów kodów programów napisanych w danym języku i wykonujących określone operacje podstawowe. Stosowanie bibliotek we własnych programach skraca czas tworzenia aplikacji oraz usprawnia jej działanie. W niektórych projektach istotny może być również warunek niezależności aplikacji od systemu operacyjnego[157].

Na potrzeby badawcze niniejszej książki napisany został program dokonujący wstępnego przetworzenia analizowanych tekstów oraz porównań lingwistycznych na podstawie przygotowanych odpowiednio tekstów. Algorytmy analizy dokumentów oraz obsługi bazy danych zostały zaimplementowane w otwartym i darmowym języku Python (wersja 2.6.5 i 2.7)[158]. Pliki zawierające wyniki poszczególnych etapów analizy dokumentów przechowywane są w odpowiednich folderach, a powiązania pomiędzy nimi zachowano przy pomocy nazw określających plik źródłowy oraz rozszerzeń wskazujących typ zawartości pliku wynikowego. Rozwiązanie takie tworzy tzw. płaską strukturę danych. W systemach przetwarzania języka naturalnego osadzonych w nurcie IR do przechowywania dodatkowych informacji o treści dokumentów stosuje się zazwyczaj bazy danych[159], dzięki czemu zoptymalizowany jest czas dostępu do danych przez wielu użytkowników równocześnie. Dużą zaletą baz danych, docenianą szczególnie w przypadku systemów wyszukiwawczych, jest szybka obsługa wielu zapytań jednocześnie (wielu operacji odczytu danych z bazy lub zapisu danych do bazy). Dane przechowywane w systemie plików oferują także możliwość równoczesnego odczytu przez wielu użytkowników, natomiast w danym momencie możliwa jest tylko jedna operacja zapisu danych do pliku. Z kolei zaletą struktury plików w porównaniu z bazą danych są mniejsze wymagania technologiczne.

---

[157] Wiele aplikacji wykorzystywanych we współczesnych projektach związanych z komputerowym przetwarzaniem języka naturalnego tworzonych w Polsce zostało napisanych w języku JAVA. Najważniejsze z nich to np. interfejs niezależnej (ang. *standalone*, działającej bez konieczności dostępu do Internetu) wersji programu POLIQARP, obsługującego pracę z Korpusem Języka Polskiego IPI PAN (por. A. Przepiórkowski, D. Janus: *POLIQUARP 1.0: Some technical aspects of a linguistic search engine for large corpora* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://nlp.ipipan.waw.pl/~adamp/Papers/2006-poliqarp/.

[158] Por. *Python License* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.python.org/psf/license/.

[159] Najczęściej są to bazy MySql lub SQLite.

Aplikacja będąca narzędziem badawczym użytym w niniejszej książce, z założenia nie obsługuje równoległych odwołań do plików czy danych. Zarówno przetwarzanie wstępne dokumentów, jak i analiza ich treści odbywają się w sposób sekwencyjny. Stworzony program komputerowy jest aplikacją przeznaczoną dla jednego użytkownika, więc imperatyw skrócenia czasu dostępu do danych nie jest w tym przypadku tak istotny. Pominięcie mechanizmów obsługi baz danych w aplikacji umożliwiło ograniczenie wielkości kodu skryptów, co przyczyniło się również do ograniczenia możliwości wystąpienia potencjalnych błędów w działaniu aplikacji.

3.2.2. Język Python[160]

Język Python należy do kategorii **języków interpretowanych**. Cechą charakterystyczną takich języków jest przetwarzanie oraz wykonywanie **kodu źródłowego** przez specjalną aplikację, tzw. interpreter[161]. Sam zaś kod źródłowy programu tworzony jest najczęściej w postaci **skryptu** zapisywanego w formie pliku tekstowego. Języki skryptowe generalnie obciążają zasoby komputera w większym stopniu niż języki kompilowane, dostarczające gotowe programy wykonalne, natomiast sam proces tworzenia i testowania aplikacji przebiega zdecydowanie szybciej w przypadku języków interpretowanych. Języki skryptowe są również w wysokim stopniu niezależne od systemu operacyjnego, ponieważ dla każdej wersji systemu dostępne są odpowiednie interpretatory, dzięki czemu jeden skrypt może działać na wielu z nich, bez potrzeby jakichkolwiek modyfikacji. Skrypty mogą być z powodzeniem uruchamiane w środowisku we-

[160] Na temat języka Python oraz możliwości jego zastosowania por. m.in.: *Python Programming Language Official Web Site* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.python.org/; *Overview Python v2.6.5 documentation* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://docs.python.org/; *The Python Wiki* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://wiki.python.org/moin/.

[161] Język interpretowany, por. hasło *interpreter*. W: *Wikipedia, Wolna encyklopedia* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/Interpreter_(program_komputerowy). Kod źródłowy jest to program komputerowy zapisany w oryginalnej postaci zgodnie ze składnią wybranego języka programowania. Przed wykonaniem kod źródłowy musi zostać skompilowany do postaci wykonalnej lub zinterpretowany przez translator, por. P. Adamczewski: *Słownik informatyczny*. Gliwice: Wydawnictwo Helion 2005, cyt. za: *WIEM, Portal Wiedzy* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://portalwiedzy.onet.pl/135705,,,,kod_zrodlowy,haslo.html; por. też hasło *kod źródłowy*. W: *Wikipedia*... [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://pl.wikipedia.org/wiki/Kod_źródłowy.

bowym, co dodatkowo zwiększa uniwersalność tego typu programów. Język Python ponadto kompiluje wykonywany kod źródłowy do postaci tzw. kodu binarnego, czyli postaci zoptymalizowanej dla danego interpretera, dzięki czemu kolejne wywołanie danego programu zajmuje mniej czasu i zasobów komputera[162]. Interpreter języka Python bezpośrednio oferuje skromną liczbę operacji podstawowych, natomiast wszystkie operacje zaawansowane przeprowadzane są za pomocą odpowiednich bibliotek podprogramów. Te dodatkowe biblioteki instalowane są razem z interpretatorem języka Python i są dostępne za pomocą polecenia *import <nazwa modułu lub biblioteki>*. Rozwiązanie takie pozwala ograniczyć znacząco wielkość pamięci operacyjnej wymaganej do pracy przez programy napisane w tym języku. Do pamięci operacyjnej wgrywane są wyłącznie potrzebne dla danej aplikacji biblioteki, dzięki czemu więcej pamięci można poświęcić na przechowywanie danych, co ma wpływ na przyspieszenie pracy programu.

O zaletach tego języka może świadczyć fakt wykorzystania go do tworzenia całości lub części interfejsów wielu portali sieciowych. W wersji webowej (jako moduł Django[163]),Python jest wykorzystywany m.in. przez wyszukiwarkę Google, serwis YouTube.com czy portal Yahoo![164].

Pliki instalacyjne interpretera języka Python można pobrać z głównej strony projektu – *Python Programming Language – Official Website,* pod adresem http://www.python.org/. Jak podają autorzy *Natural lan-*

---

[162] Por. *Using Python. Release 2.6.5*, pod red. G. Rossum van, F. L. Drake Jr. W: *Using Python – Python v2.6.5 documentation* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://docs.python.org/using/index.html. Dokument ten, jak i wiele innych związanych z wersją 2.6.5 języka Python, dostępny jest w witrynie *Python v2.6.5 documentation* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://docs.python.org/index.html oraz w postaci pakietu zip na stronie *Download Python v2.6.5 documentation* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://docs.python.org/download.html.

[163] Na temat modułu Django oraz możliwości jego zastosowania por. *Django. The Web framework for perfectionists with deadlines* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.djangoproject.com/; *Django. Django documentation*. [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://docs.djangoproject.com/en/1.2/; *Django. Framework webowy dla perfekcjonistów z terminami* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.django.pl/.

[164] Pełna lista projektów korzystających z języka Python znajduje się na stronie *Python Success Stories* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.python.org/about/success/ oraz w postaci opinii użytkowników na stronie *Quotes about Python* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.python.org/about/quotes/.

guage processing with Python, język ten cechuje się przejrzystą składnią i budową oraz, co jest istotne dla niniejszej książki, dobrze obsługuje dane typu **łańcuchowego**[165]. Kolejną zaletą tego języka, wymienianą m.in. w przytaczanym podręczniku, jest jego obiektowość – każda zmienna posiada zdefiniowane atrybuty i metody. Jako język obiektowy Python oferuje możliwość łatwego wykorzystania danych oraz ich metod, natomiast jako język dynamiczny umożliwia elastyczne modyfikowanie właściwości obiektów, np. dodawanie nowych metod[166].

3.3. Organizacja i przebieg badań

W celu zrealizowania zamierzeń badawczych przygotowano kolekcję dokumentów, których teksty były obiektem dalszych operacji i analiz.

Na potrzeby ustalenia skuteczności automatycznego wskazywania słów kluczowych zostały przeprowadzone badania porównawcze. Autor niniejszej książki stworzył aplikację komputerową – zestaw skryptów języka Python wykonujących operacje wstępnego przetworzenia tekstu dokumentu oraz analizujących uzyskane teksty. Docelowo skrypty mają wskazać słowa kluczowe związane z treścią analizowanych dokumentów. Dodatkowo analizowane były dokumenty posiadające słowa kluczowe ustalone przez swych autorów oraz przez niezależne osoby. Procesowi badawczemu poddawane były pojedyncze artykuły będące fragmentami prac zbiorowych, zarówno wydawnictw regularnych (czasopism), jak i materiałów konferencyjnych[167].

W kolejnych podrozdziałach zostały zaprezentowane kolejne etapy prac badawczych oraz zasady ich przeprowadzenia.

3.3.1. Przygotowanie dokumentów do analizy

W celu przygotowania treści dokumentów do dalszego opracowania i analizy konieczne było zapisanie treści artykułów w postaci plików tekstowych. Wspólny format i sposób zapisu danych badawczych podyktowany jest wymogiem usprawniającym jakiekolwiek automatyczne operacje na tekstach dokumentów.

---

[165] Dane typu łańcuchowego są to dane tekstowe w dowolnej skali, tzn. całe teksty, poszczególne wyrazy, jak i poszczególne litery w pojedynczych wyrazach.

[166] Por. S. Bird, E. Klein, E. Loper: *Natural language processing with Python. Analyzing text with the Natural Language Toolkit*. Sebastopol: O'Reily 2009, s. XII-XIII.

[167] Na potrzeby opisu organizacji i przebiegu badań zamiennie stosowane będą terminy artykuł oraz tekst.

Większość dokumentów poddanych analizie w trakcie badań jest dostępna oryginalnie w postaci plików DjVu lub PDF. Z powodu dużego bogactwa wewnętrznych znaczników formatu automatyczna analiza treści dokumentu jest utrudniona, przy czym możliwość kodowania znaków języka polskiego według kilku standardów[168] stanowi dodatkowy problem dla procesu analizy.

Bezpośrednie wczytanie zawartości pliku PDF przez skrypt języka Python dało rezultaty zobrazowane na ilustracji 2. Został wczytany kod binarny .PDF, w którym tekst artykułu wyświetlany jest łącznie ze znacznikami kontrolnymi. W takiej postaci tekst treści jest niedostępny, nie można wykonać na nim żadnych operacji – czy to automatycznych, czy przeprowadzonych przez człowieka.

Ilustracja 2. Podgląd zawartości przykładowego pliku .pdf wczytana bezpośrednio przez skrypt języka Python.

Źródło: Opracowanie własne na podstawie dokumentów zebranych w celach analizy.

Język Python posiada co prawda wiele bibliotek związanych z przetwarzaniem dokumentów pdf, jak np. *pypdf*[169], czy dających się zastosować przez skrypty tego języka pakietów oprogramowania typu *xpdf*[170], jednak-

[168] Najpopularniejszymi ze standardów kodowania są: środkowoeuropejski ISO88592, środkowoeuropejski Windows1250 czy UTF8. Każdy z nich w odmienny sposób zapisuje narodowe znaki diakrytyczne, stąd przy zastosowaniu nieprawidłowego formatu do odczytu zawartości pliku można uzyskać niepoprawne litery w tekstach.

[169] Por. *pyPdf* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://pybrary.net/pyPdf/.

[170] Por. *Xpdf* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.foolabs.com/xpdf/.

że błędnie interpretują one polskie znaki, tzw. diakrytyczne, czyli litery: *ą, ć, ę, ł, ń, ó, ś, ź, ż*[171]. W związku z tym dokumenty w formacie .pdf pozyskane do analizy zostały przekonwertowane przy użyciu dodatkowych aplikacji do postaci tekstowej[172].

Podobnie postąpiono z plikami źródłowymi w formacie djvu. Tak jak format .pdf, nie poddaje się on bezpośredniemu odczytowi przez aplikacje języka Python. Dlatego też również pliki .djvu konwertowane były do postaci plików tekstowych.

Wszystkie etapy automatycznej analizy treści dokumentów przeprowadzane były na plikach źródłowych zapisanych w formacie .txt. Jako wspólny standard kodowania znaków wybrano UTF8[173].

Ze względu na postawione hipotezy badawcze, szczególnie tę związaną z wpływem wyrażeń wyróżnionych na wagę danego słowa, w plikach tekstowych zostały wskazane wybrane elementy mogące mieć wpływ na generowanie automatycznie słów kluczowych. Były to wyróżnione przez autorów artykułów frazy i wyrażenia. Oprócz nich wyróżniono również elementy identyfikacyjne dla każdego pliku, takie jak: autor, tytuł i autorskie słowa kluczowe. Operacje oznaczania elementów informacyjnych przeprowadzone były metodą tradycyjną.

Poszczególne elementy treściowe w plikach informacyjnych zostały opatrzone odpowiednimi znacznikami definiującymi znaczenie zapisanych danych. Idea znaczników została zaczerpnięta ze standardu HTML, podobny sposób znakowania wykorzystywany jest również przez innych badaczy języka naturalnego[174]. Każdy użyty znacznik posiada znak zamy-

---

[171] W miejscu występowania wymienionych znaków testowane aplikacje rozdzielały wyraz polski na odrębne części wstawiając symbol spacji zamiast litery polskiej.

[172] Ostatecznie, z powodu problemów, jakie różnym aplikacjom konwertującym zapis .pdf do formatu tekstowego sprawiało prawidłowe kodowanie polskich znaków, zdecydowano się na rozwiązanie pośrednie. Oryginalne pliki .pdf zostały początkowo przekonwertowane do plików typu .djvu za pomocą nieodpłatnej aplikacji Pdf To Djvu GUI v.2.1, ze strony *Pdf to DjVu GUI main page (official page)* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.trustfm.net/GeneralTools/SoftwarePdfToDjvuGUI.php. Konwersja ta zachowywała prawidłową pisownię wyrazów z tekstu. Następnie pliki .djvu były konwertowane na pliki tekstowe (.txt) za pomocą nieodpłatnej aplikacji DjVuLibre dostępnej na stronie *GB Soft – archiwizacja, udostępnianie dokumentacji. Oprogramowanie DjVu, djvu viewer.* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.djvu.com.pl/download.php.

[173] Każdy wynikowy plik tekstowy powstały na etapie konwersji formatu zapisu do postaci .txt był kodowany w standardzie UTF8.

[174] Por. np. tagowanie pozycji w korpusie PAN.

kający, składający się z nazwy znacznika poprzedzonej symbolem (/)[175]. Tabela 5. prezentuje znaczniki zastosowane w celu wyróżnienia elementów informacyjnych oraz ich znaczenie.

Tabela 5. Znaczniki zastosowane w plikach metainformacyjnych o analizowanych tekstach oraz ich znaczenie.

Znacznik	Znaczenie
<id>	identyfikator tekstu*
<plk>	ścieżka dostępu do pliku z treścią dokumentu
<w>	liczba wyrazów w tekście właściwym dokumentu
<s>	liczba leksemów w tekście właściwym dokumentu
<au>	autorzy
<t>	tytuł
<abs>	abstrakt
<sk>	słowa kluczowe autorskie

* Ze względów praktycznych nazewnictwo plików na potrzeby niniejszej książki zostało ustandaryzowane. W dalszej części analiz automatycznych odwołania do poszczególnych plików odbywają się zawsze za pośrednictwem identyfikatorów (wartości pól <ID>).

Źródło: Opracowanie własne.

Pola *autor* oraz *słowa kluczowe* mogą zawierać więcej niż jedną pozycję, w takim przypadku kolejne wartości oddzielane są od siebie znakiem średnika (;)[176].

Pliki informacyjne wykorzystywane są do analizy plików z treścią, jak i do przechowywania danych opisujących treść, natomiast właściwa treść dokumentów przechowywana jest w odrębnych plikach tekstowych identyfikowanych przez wartości znaczników <ID> oraz pośrednio przez <PLK>. Nazwy plików wejściowych składają się z numeru (wyrażonego za pomocą liczb arabskich przypisywanych w kolejności rosnącej) oraz literowego identyfikatora kolejnego artykułu w kolekcji (litery alfabetu przypisywane w kolejności rosnącej).

[175] Znacznik zamykający informuje program analizujący o końcu poszczególnych pól informacyjnych lub o końcu danego sposobu formatowania w przypadku treści dokumentów.
[176] Średnik jest standardowym znakiem delimitacji tekstu używanym m.in. w standardzie csv.

3.3.2. Klasyfikacja zawartości pliku

Ponieważ celem książki było rozpoznanie możliwości generowania słów kluczowych i porównanie procesu ich wskazywania automatycznego z dokonanym przez człowieka, do analizy wybrane zostały teksty zaopatrzone przez swych autorów w zestawy słów kluczowych charakteryzujących treść dokumentów. Analizie poddano artykuły pochodzące z prac zbiorowych. Pojedyncze artykuły zostały wyodrębnione z dokumentów źródłowych i zapisane w postaci indywidualnych plików, jak zostało to zaprezentowane w poprzednim podrozdziale. Każdy dokument na etapie konwersji był dzielony na część informacyjną (metainformacje) oraz część treściową. Przykładowe pliki związane z jednym z analizowanych dokumentów prezentuje ilustracja 3.

Ilustracja 3. Lista plików związanych z przykładowym analizowanym dokumentem.

Nazwa	Data modyfikacji	Typ	Rozmiar
1a.bow	2010-11-30 16:52	Plik BOW	11 KB
1a.frek	2010-10-28 13:30	Plik FREK	8 KB
1a.lem	2010-10-28 13:31	Plik LEM	11 KB
1a.lemfrk	2010-10-28 13:32	Plik LEMFRK	5 KB
1a.met	2010-10-29 13:01	Plik MET	2 KB
1a.txt	2010-10-28 13:28	Dokument tekstowy	18 KB
1a.wyr	2010-10-28 13:30	Plik WYR	1 KB

Źródło: Opracowanie własne na podstawie plików przygotowanych do analizy.

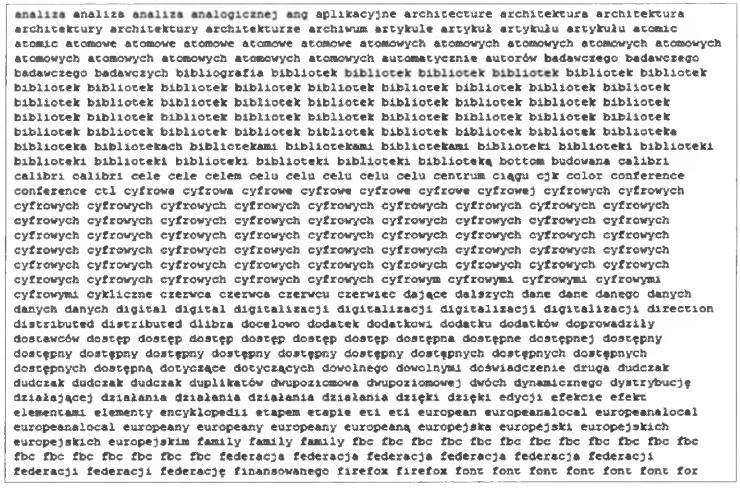
Zastosowane rozszerzenia w nazwach plików oznaczają odpowiednio:

.bow – plik typu *bag-of-words* (wielozbiór), plik przechowujący wszystkie wyrazy z tekstu artykułu pozostałe po usunięciu słów nieznaczących. Wyrazy zapisane są w formie, w jakiej wystąpiły w artykule, w porządku alfabetycznym. Pliki typu .bow odzwierciedlają wyłącznie zawartość wyrazową tekstu, a pomijają kolejność wystąpienia i inne informacje pozatekstowe. Zawartość przykładowego pliku prezentuje ilustracja 4.

.frek – plik prezentujący wyrazy z pliku typu .bow posortowane w kolejności malejącej według liczby wystąpień w dokumencie z podanymi frekwencjami, jak na ilustracji 5.

.lem – plik zawierający wyrazy pozostałe po operacji optymalizacji (zapisane w pliku .bow) sprowadzone do postaci podstawowej, tzw. lematy. Przykładową zawartość prezentuje ilustracja 6.

Ilustracja 4. Zawartość przykładowego pliku typu .bow.



Źródło: Opracowanie własne na podstawie plików przygotowanych do analizy.

Ilustracja 5. Zawartość przykładowego pliku typu .frek posortowana według liczby powtórzeń.

biblioteka	28
biblioteki	20
zbiorów	14
cyfrowej	14
cyfrowa	14
bibliotek	14
politechniki	12
podlaska	12
podlaskiej	11
liczba	10
konsorcjum	10
białostockiej	9
warszawa	8
rysunek	8
publikacji	8
dokumentów	7
digitalizacji	7
bibliotecznych	7
białymstoku	7
źródło	6
pbc	6
dygitalizacja	6

Źródło: Opracowanie własne na podstawie plików przygotowanych do analizy.

Ilustracja 6. Zawartość pliku przechowującego wyrazy z dokumentu w postaci lematów.

```
access acrobat adekwatny adobe adres afisz akademicki akademicki akademia akademia
akademia analiza analiza ankieta anna anna anna aparat aparat archidiecezjalny archiwizacja
archiwum artykuł autor autor autor autorski autorski autorski autorski autorski autor
autor barbara barbara barbara baza bazować baza bezpośredni bezzubik bezzubik Białystok
białostocki białostocki białostocki białostocki białostocki białostocki białostocki
białostocki białostocki białostocki białostocki Białystok Białystok Białystok Białystok
Białystok Białystok Białystok białystok bibliografia biblioteka biblioteczny biblioteczny
biblioteczny biblioteczny biblioteczny biblioteczny biblioteczny biblioteka biblioteka
biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka
biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka
biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka
biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka
biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka
biblioteka biblioteka bibliotekarz biblioteka biblioteka biblioteka biblioteka
biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka
biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka biblioteka
biblioteka biblioteka biuletyn biuletyn biznes bogiel bogiel bogiel bottom bpb brać budować
budynek błąd cebid cel cel cel cel cenne centralny centralny centrum Chopin chronić
cieszyć ciąg coś content copyright copyright core cyfrowy cyfrowy cyfrowy cyfrowy
cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy
cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy
cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy
cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy cyfrowy czasochłonny czasowo czcionka czerpać
Czerwiec Czerwiec Czerwiec czterech czytelnik czytelnik czytelnik czytelnik częsty
częstotliwość częsty część członek dać dedykować digitalizacja digitalizacja digitalizacja
digitalizacja digitalizacja digitalizacja digitalizacja digitalizacja digitalizacja
digitalizacja digitalizacyjnej digitalizowanych digitalizować dlibra dlibra dlibra dlibra dlibra
```

Źródło: Opracowanie własne na podstawie plików przygotowanych do analizy.

.lemfrk – plik zawierający lematy posortowane w kolejności malejącej według liczby wystąpień danego słowa, jak na ilustracji 7.

Ilustracja 7. Lematy posortowane malejąco według liczby wystąpień.

```
biblioteka 67
cyfrowy 40
podlaski 25
zbiór 19
liczba 13
politechnika 12
naukowy 12
dokument 11
białostocki 11
publikacja 10
konsorcjum 10
digitalizacja 10
warszawa 9
kolekcja 9
rysunek 8
dostępny 8
Białystok 8
źródło 7
zasób 7
dygitalizacja 7
biblioteczny 7
wydawnictwo 6
strona 6
prawo 6
praca 6
pbc 6
internet 6
```

Źródło: Opracowanie własne na podstawie plików przygotowanych do analizy.

.met – plik zawierający metainformacje dotyczące analizowanego dokumentu. Pliki .met zostały opisane w dalszej części rozdziału, natomiast zawartość przykładowego pliku .met prezentuje ilustracja 8.

Ilustracja 8. Zawartość przykładowego pliku .met.

```
<id>lz.txt</id>
<plik>"D:\korpusPM\dane\InfNauk\lz.txt"</plik>
<w> 1590 </w>
<s> 591 </s>
<au>Barbara Kubiak, Anna Bogiel-Kuźmicka</au>
<t>Podlaska Biblioteka Cyfrowa </t>
<abs>
Celem artykułu i jest przedstawienie dorobku Podlaskiej Biblioteki Cyfrowej istniejącej od 2004
roku, współtworzonej przez biblioteki miasta Białegostoku. Omówiono  zadania:
   - rozpowszechnianie dorobku naukowego pracowników poszczególnych uczelni;
   - ochrona najcenniejszych zbiorów;
   - zdalne udostępnianie zbiorów bibliotecznych.
Zaprezentowano zbiory oraz statystyki ich wykorzystania. Przedstawiono również problemy
z udostępnianiem zbiorów. </abs>
<slk>biblioteka cyfrowa; digitalizacja</slk>
```

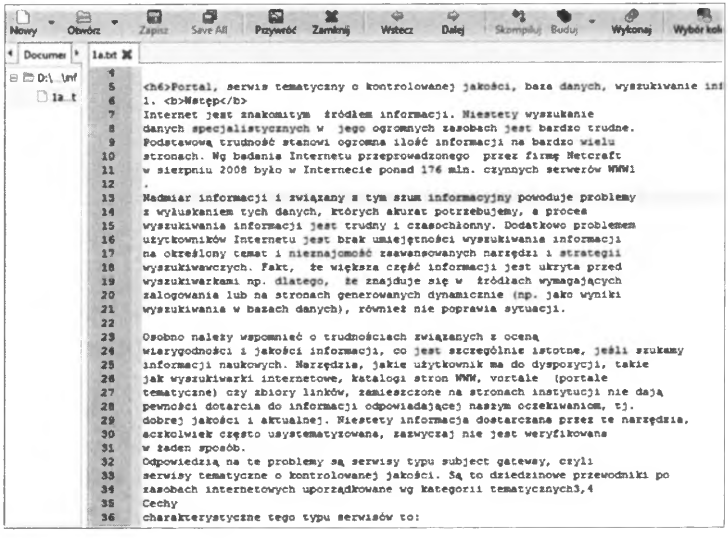
Źródło: Opracowanie własne na podstawie plików przygotowanych do analizy.

.txt – plik w formacie tekstowym zawierający treść i wybrane elementy formatowania wyrażeń. Zawartość plików .txt jest podstawą dalszych operacji przetwarzania i analizy treści dokumentów. Standardowo na potrzeby badań nad tekstami języka naturalnego pomijane są całkowicie informacje o formatowaniu tekstu, badacze koncentrują się na frekwencjach wyrazów oraz rozpoznawaniu poszczególnych słów i relacji pomiędzy nimi. W bieżącej pracy natomiast skoncentrowano się na słowach kluczowych oraz wyróżnionych przez autorów dokumentów wyrażeniach, co spowodowało konieczność zachowania formatowania części tekstów. Słowa kluczowe wyznaczane są na podstawie tytułu i śródtytułów dokumentu, jego treści oraz w uzasadnionych przypadkach również na podstawie cech pozatreściowych. Powinny one jak najprecyzyjniej opisywać treść dokumentu. Oprócz tytułów i śródtytułów pomocne przy ustalaniu słów kluczowych mogą być wyrazy i wyrażenia wyróżnione przez autorów w tekście dokumentu.

W związku z powyższym w analizowanych publikacjach zachowane zostały informacje o wyróżnieniach wyrazów lub ich ciągów w poszczególnych dokumentach. Formatami uwzględnianymi w trakcie analizy były: wyróżnienia nagłówków 1, 2 i 3 poziomu oraz wyróżnienia typu kursywa, podkreślenie lub pogrubienie czcionki. Z powodów praktycznych – w celu ułatwienia przetwarzania wybranych fragmentów tekstu – wszystkie wymienione znaczniki zostały w plikach roboczych przedstawione w jednolitej formie. Do wskazania wyróżnionych fragmentów tekstu przyjęto znacznik <b> wykorzystywany standardowo do oznaczenia wytłusz-

czenia tekstu w dokumentach HTML. Zawartość przykładowego pliku .txt została zaprezentowana na ilustracji 9.

Ilustracja 9. Zawartość przykładowego pliku .txt zawierającego treść analizowanego dokumentu.



Źródło: Opracowanie własne na podstawie plików przygotowanych do analizy.

.wyr – plik zawierający słowa wyróżnione w tekście odpowiednim formatowaniem. Przykładowy zestaw wyróżnionych elementów treści dokumentu prezentuje ilustracja 10.

Ilustracja 10. Wyrażenia wyróżnione w tekście analizowanego artykułu.

```
Geneza Podlaskiej Biblioteki Cyfrowej
Cele i zasoby Podlaskiej Cyfrowej
Rozwój Podlaskiej Biblioteki Cyfrowej
Podsumowanie
Bibliografia
```

Źródło: Opracowanie własne na podstawie plików przygotowanych do analizy.

3.3.3. Usunięcie wyrazów nierelewantnych

Przygotowana wstępnie treść dokumentu zostaje poddana operacji zamiany wszystkich wielkich liter na małe, co ma znaczenie przy automatycznym wyszukiwaniu i przetwarzaniu wyrazów przez język Python. Ko-

lejnym etapem wstępnego przetwarzania dokumentu jest usunięcie wyrazów nieznaczących. Jako takie traktowane są:

- znaki nienumeryczne, np. „&", „^", „%", „-",
- liczby,
- wyrazy jedno- oraz dwuliterowe, jak np. a, i, z, w, o, na, we, ze, za, ponieważ nie wnoszą one wartościowej informacji do tekstu,
- wyrazy z listy słów nieznaczących, która została zaprezentowana w rozdziale *Ustalenia terminologiczne oraz wybrane metody komputerowego przetwarzania języka naturalnego*.

Z zawartości pliku przefiltrowanej zgodnie z wymienionymi powyżej kryteriami powstaje nowy tekst, zawierający wyłącznie słowa znaczące o długości powyżej dwóch liter[177]. Jest to przykład przechowywania zoptymalizowanej treści dokumentu w postaci wielozbioru (ang. *bag-of-words*). Dodatkowo tekst wynikowy, przed zapisaniem w nowym pliku, zostaje posortowany alfabetycznie. W wyniku tej operacji dla każdego analizowanego dokumentu uzyskuje się nowy plik zawierający zoptymalizowany tekst.

3.3.4. Ustalenie podstawowej postaci wyrazów

Kolejną operacją, której poddawana jest zoptymalizowana treść dokumentu, jest sprowadzenie poszczególnych wyrazów do formy hasłowej. Jak już wspomniano w rozdziale 2., czynność ta została wykonana na podstawie porównania poszczególnych wyrazów z przefiltrowanego tekstu do list wzorcowych zawierających poprawne formy poszczególnych słów języka polskiego. Porównanie dokonywane jest na zasadzie podobieństwa ciągu znaków (liter), a nie znaczenia. W wyniku porównania wszystkie jednostki tekstowe zawierające końcówki fleksyjne zostały zastąpione odpowiednimi leksemami hasłowymi.

Procedurę lematyzacji przeprowadzono po wyłączeniu z treści dokumentu słów nieznaczących. Z dwóch dostępnych możliwości (algorytm oraz metoda słownikowa) wybrano metodę słownikową. Pozwala ona wskazać formę podstawową dla wszystkich prawidłowych odmian danego wyrazu zapisanych w odpowiednim słowniku. Ponieważ w niniejszej pracy nie badano związków pomiędzy wyrazami w tekstach oraz usuwano wyrazy nieznaczące, rozwiązanie wykorzystujące gotowy słownik odmian było w pełni wystarczające.

[177] Na wyłączenie z dalszych analiz wyrazów o długości do 2 znaków zdecydowano się ze względu na obniżenie kosztów i skrócenie czasu wykonywania operacji na tekstach. Za decyzją taką przemawia również niska wartość informacyjna tak krótkich słów.

Operacja lematyzacji w omawianym systemie przeprowadzana jest na podstawie dostępnego w ramach licencji GPL *Słownika odmian*, notującego 180.140 form podstawowych wyrazów oraz prawidłowych postaci odmian poszczególnych słów[178]. Należy przyznać, że wartość ta jest niemała, ponieważ np. wydawcy *Wielkiego słownika ortograficznego języka polskiego* informują, że ich słownik zawiera ok. 140.000 haseł[179]. Operacja sprowadzania wyrazów do wspólnej postaci wymaga porównania słowoformy występującej w treści z zawartością słownika i pobraniu wyrazu w postaci podstawowej z pierwszej pozycji wiersza, w którym znaleziono dane wystąpienie. Słowoforma z treści dokumentu zostaje następnie zamieniona na lemat pobrany z pliku słownika. Jednakże, jak zazwyczaj zdarza się w przypadku korzystania z jakiegokolwiek słownika, nie wszystkie wyrazy występujące w analizowanych dokumentach są przez słownik odmian rejestrowane. W takich przypadkach wyraz zostaje w procedurze lematyzacji zachowany w niezmienionej formie, a następnie poprawiony przez operatora, przy czym do odpowiedniego słownika zostaje dopisany odpowiedni wiersz z odmianami, w celu zautomatyzowania kolejnych przekształceń w przyszłości.

Odrębnym problemem jest nieprawidłowe przypisanie formy podstawowej dla danego wyrazu w sytuacji homonimii. Metody słownikowe zazwyczaj przyjmują dla homonimów formę podstawową pierwszego w kolejności alfabetycznej znaczenia. Czasami forma podstawowa danego wyrazu występująca w tekście dokumentu była zmieniana przez program na formę podstawową innego leksemu z powodu tożsamości zapisu z wcześniejszą alfabetycznie formą odmiany innego leksemu. Sytuacje takie, jeśli zostały wykryte w trakcie osobiście przeprowadzonego procesu weryfikacji, były korygowane. Jednakże, należy nadmienić, że nieprawidłowe przypisania dotyczyły wyrazów o niskich frekwencjach wystąpień, zatem nieistotnych dla dalszych procesów analizy.

Słownik odmian prezentuje posortowaną alfabetycznie listę słowoform, ich lematów oraz prawidłowych postaci gramatycznych danego słowa. Informacje dotyczące danego słowa zapisane są w jednym wierszu, a całość zapisana w postaci pliku tekstowego. Ze względu na rozmiar pliku oryginalnego (52 MB) oraz specyfikę procesu porównawczego

[178] Opracowanie własne na podstawie zawartości *Słownika odmian*.

[179] Cyt za: *Poradnia językowa* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://poradnia.pwn.pl/lista.php?id=5606.

wykonywanego przez program[180], zawartość słownika została na potrzeby niniejszej książki podzielona na sekcje zapisane w odrębnych plikach. W związku z powyższym oryginalny plik został podzielony na pliki zawierające wyrazy zaczynające się na poszczególne litery alfabetu, a sama nazwa pliku odzwierciedla jegozawartość, np. *a.txt* zawiera fragment słownika dla wyrazów zaczynających się na literę *a*. Krok ten został podyktowany ograniczeniami technicznymi. Takie podejście połączone z sortowaniem wyrazów w plikach .bow pozwala przeprowadzać operację lematyzacji dla ograniczonego zbioru wyrazów zaczynających się na jedną literę, a tym samym odwoływać się do jednego tylko pliku słownikowego. Oczywiście, proces ten jest powtarzany dla wszystkich początkowych liter wyrazów występujących w danym pliku.

Podział słownika na funkcjonalne, niezależne części, w sposób ułatwiający zlokalizowanie konkretnego poszukiwanego elementu (tu wyrazu z treści analizowanego dokumentu), jest zbliżone do mechanizmu **haszowania**. W pracy *Komunikacja człowieka z maszyną. Komputerowe modelowanie kompetencji językowej* Zygmunt Vetulani definiuje funkcję haszującą jako funkcję przyporządkowującą słowu hasłowemu (kluczowi) pewien adres, za pomocą którego możliwy jest szybki dostęp do danego klucza. Zaletą dobrze zaprojektowanej operacji haszowania jest zapewnienie szybkiego i niezależnego od wielkości słownika dostępu do haseł[181].

W przytaczanej pracy Z. Vetulani zauważa, że słownik form języka polskiego jest na ogół nadmiarowy w stosunku do analizowanego tekstu (wszystkie poprawne formy danego słowa w porównaniu do kilku form występujących w dokumencie). Jednym ze sposobów ograniczenia wpływu tej nadmiarowości na efektywność działania aplikacji jest odpowiednio zaprojektowana funkcja haszująca[182].

Zawartość fragmentu pliku słownikowego wykorzystywanego w niniejszej książce prezentuje ilustracja 11.

---

[180] Procedura porównania wymagałaby przeszukania wyrazów występujących w porządku alfabetycznym przed badanym słowem, co znacznie wydłuża czas operacji.

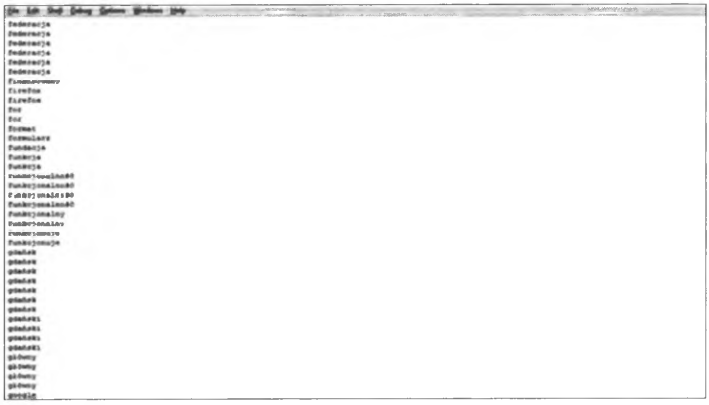
[181] O zastosowaniu funkcji haszujących w przetwarzaniu tekstów języka naturalnego por. m.in. Z. Vetulani: *Komunikacja człowieka z maszyną. Komputerowe modelowanie kompetencji językowej.* Warszawa: Akademicka Oficyna Wydawnicza EXIT 2004, s. 132.

[182] Por. tamże, s. 146.

Ilustracja 11. Zawartość fragmentu pliku słownikowego.

Źródło: Opracowanie własne na podstawie zawartości pliku słownikowego.

Ilustracja 12. Zoptymalizowana treść dokumentu poddana procesowi lematyzacji.



Źródło: Opracowanie własne na podstawie analizy przykładowego dokumentu.

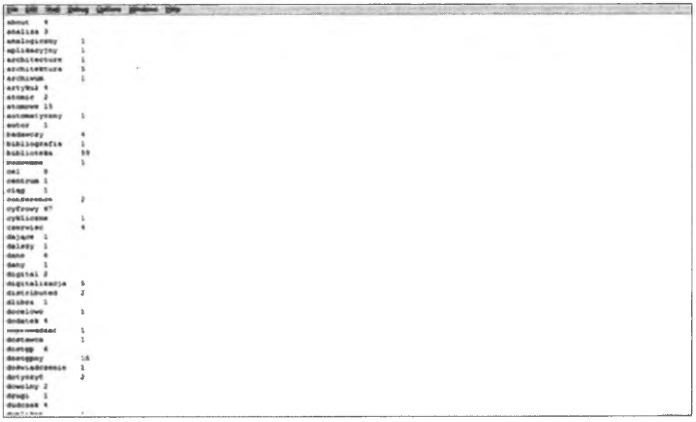
Dzięki takiej organizacji danych porównawczych operacja sprowadzania do postaci podstawowej została zoptymalizowana. Po pobraniu słowa z dokumentu system umieszcza w pamięci operacyjnej plik słownikowy, którego nazwa zaczyna się na tę samą literę, co pobrane słowo. Zawartości obu plików przechowywane są w odpowiednich zmiennych. Wynik działania operacji lematyzacji obrazuje ilustracja 12.

Po wykonaniu wymienionych operacji optymalizacji i wstępnego przetworzenia treści dokumentu następuje proces właściwej analizy tekstu, który również składa się z kilku etapów. Pierwszym z nich jest analiza frekwencji poszczególnych leksemów w dokumencie.

3.3.5. Zliczenie wystąpień danego słowa

Analiza frekwencji jest prostym policzeniem wystąpień leksemu w treści dokumentu. Przykładowy wynik tej operacji widoczny jest na ilustracji 13.

Ilustracja 13. Wynik sumowania wystąpień poszczególnych słów w danym dokumencie.



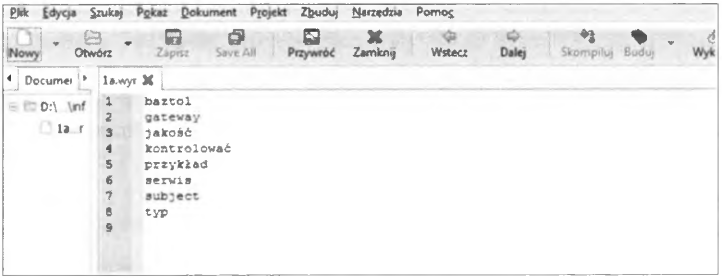
Źródło: Opracowanie własne na podstawie analizy przykładowego dokumentu.

3.3.6. Analiza słów wyróżnionych

Oprócz operacji przetwarzania zawartości pliku typowych dla badań statystycznych nad językiem naturalnym w niniejszej książce analizowane są dodatkowo słowa wyróżnione przez autorów w tre-

ści dokumentów. Jak wspomniano wcześniej, wyróżniono je znacznikiem <b>. Kolejnym, dodatkowym etapem analizy treści dokumentu jest zatem wyodrębnienie owych wyróżnionych wyrazów. Są one przechowywane w osobnym pliku, skojarzonym za pomocą nazwy z plikiem zawierającym treść dokumentu. Zawartość pliku informacyjnego o wyróżnieniach w treści poddawana jest takim samym transformacjom, jak treść zasadnicza, czyli (dla przypomnienia): zamianie liter wielkich na małe, usunięciu wyrazów nieznaczących, ustaleniu wspólnej postaci danego słowa i wreszcie zliczeniu wystąpień danego słowa. Zawartość przykładowego pliku przechowującego zoptymalizowane wyrażenia wyróżnione w tekście dokumentu przez jego autora prezentuje ilustracja 14.

Ilustracja 14. Wyrażenia wyróżnione w tekście dokumentu po wstępnym zoptymalizowaniu.



Źródło: Opracowanie własne na podstawie analizy przykładowego dokumentu.

3.3.7. Metody ustalania wagi słów

Warto przypomnieć, że słowa kluczowe są wyrażeniami charakteryzującymi zawartość treściową dokumentu. W związku z tym muszą to być słowa silnie nacechowane informacją, w dodatku informacją związaną w jak największym stopniu z treścią opisywanego tekstu. Zwykłe wskazanie częstości wystąpień danego słowa w dokumencie jest co prawda proste, ale nie odzwierciedla jego faktycznej wartości informacyjnej. Do wad tego podejścia można zaliczyć wyznaczanie wysokich wag słowom powszechnym, które powtarzają się w wielu tekstach. Ponadto słowa o najwyższych częstościach w tekstach danego języka, właśnie ze względu na niską wartość informacyjną,

często umieszczane są na stoplistach. Jednakże dla dalszych operacji analizy statystycznej częstości wystąpień są danymi podstawowymi. W literaturze przedmiotu wskazywanie częstości wystąpień oznaczane jest notacją:

$$tf_{t,d} = ?t$$

gdzie:
t – a słowo[183],
f – częstość wystąpienia,
d – dokument,
a cały zapis oznacza częstość wystąpienia słowa *t* w dokumencie *d*[184].

W związku z niską przydatnością wskazania samej częstości wystąpienia danego słowa, dla określenia jego wagi stosuje się dodatkowe porównania statystyczne. Jednym z nich jest wskazanie tzw. odwróconej częstości w dokumentach (ang. *inversed document frequency*, idf). Jest to zlogarytmowana miara liczby dokumentów z kolekcji, w których dane słowo wystąpiło, oznaczana następująco:

$$idf_t = log_2 \frac{N}{df_t}$$

gdzie:
idf – odwrócona częstość wystąpienia słowa w dokumentach z kolekcji,
N – liczba wszystkich dokumentów w kolekcji,
$\mathbf{df_t}$ – liczba dokumentów, w których dane słowo wystąpiło.

Z właściwości logarytmów wynika, że wartość idf_t będzie wysoka dla słów rzadkich, natomiast dla słów występujących w wielu dokumentach będzie ona niska, jeżeli d = N, to $\log_2 1 = 0$. Należy tu nadmienić, że w tym przypadku nie analizuje się ogólnej liczby wystąpień danego słowa w dokumencie, ale sam fakt takiego wystąpienia lub jego braku[185].

[183] Należy przy okazji zaznaczyć, że wszystkie prace anglojęzyczne konsekwentnie stosują pojęcie *term* na oznaczenie analizowanych słów.
[184] Por. m.in. Ch. D. Manning: dz. cyt., s. 118 i nast.
[185] Por. tamże, s. 119.

Obie powyższe miary wykorzystywane są w tzw. ważeniu tf-idf[186], przedstawianym wzorem ogólnym[187]:

$$tf\text{-}idf_{t,d} = tf_{t,d} idf_{t,d}$$

Wartości powyższego wyrażenia są najwyższe dla słów występujących z dużą częstością w niewielkiej liczbie dokumentów, osiągają z kolei wartości niższe dla słów pojawiających się rzadko w dokumentach oraz najniższe dla słów pojawiających się w większości dokumentów z kolekcji. Wyniki tego porównania są zgodne z intuicyjnym wyznaczaniem wagi danego słowa.

Metoda ważenia słów td-idf jest stosowana głównie w systemach informacyjno-wyszukiwawczych (jak np. wyszukiwarki internetowe, systemy klasyfikujące) dla wskazania dokumentów podobnych. Inne możliwe sposoby wskazywania wagi słowa w dokumencie uwzględnione w niniejszej książce to porównanie częstości wystąpienia danego słowa do liczby wystąpień najpowszechniejszego słowa w dokumencie oraz porównanie frekwencji słowa w danym dokumencie do ogólnej frekwencji tego samego słowa w całej kolekcji analizowanych dokumentów.

3.3.8. Porównanie zestawów słów kluczowych ustalanych tradycyjnie i automatycznie

Ostatni etap badań polegał na porównaniu zestawów słów kluczowych wybranych w wyniku procesów kognitywnych (przez autorów poszczególnych dokumentów oraz przez indeksatorów) i wygenerowanych automatycznie. Słowa kluczowe ustalane przez autorów analizowanych tekstów były pobierane automatycznie z odpowiednich plików informa-

[186] O ustalaniu wagi słów metodą tf-idf por. K. Spärck Jones*: A statistical interpretation of term specificity and its application in retrieval* [on-line]. [Dostęp: 13 listopada 2011]. Dostępny w World Wide Web: http://www.soi.city.ac.uk/~ser/idfpapers/ksj_orig.pdf – w pracy tej autorka wprowadza pojęcie wagi termu i przeprowadza rozważania na temat możliwych sposobów ustalania tej wagi. Z kolei w pracy, która jest swego rodzaju repliką na referowany artykuł (Stephen Robertson*: Understanding inverse document frequency: on theoretical arguments for IDF* [on-line]. [Dostęp: 13 listopada 2011]. Dostępny w World Wide Web: http://www.soi.city.ac.uk/~ser/idfpapers/Robertson_idf_JDoc.pdf) autor dyskutuje niezależność uzyskanych wag słów od podstawy logarytmu używanego we wzorze tf-idf.

[187] Por. tamże, s. 120 i nast. Pokaźna liczba dokumentów poświęconych mierzeniu wartości słów dostępna jest na stronie *The IDF page* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.soi.city.ac.uk/~ser/idf.html.

cyjnych, słowa kluczowe ustalane przez indeksatorów były wynikiem badań przeprowadzonych w grupie studentów UMK (co zostało opisane w dalszej części rozdziału). Automatyczne generowanie słów kluczowych odbywało się na podstawie wyznaczenia wagi leksemów powstałych w wyniku opracowania lingwistycznego analizowanych tekstów. Charakterystyki poszczególnych zbiorów słów kluczowych zostały przedstawione w podrozdziale poświęconym prezentacji materiału badawczego.

3.4. Prezentacja materiału badawczego

Równolegle z przetwarzaniem lingwistycznym poszczególnych artykułów następowało budowanie korpusu tekstów z danej dziedziny. Korpus tworzony był poprzez poddanie tekstów wytypowanych dokumentów takiemu samemu procesowi przygotowania, jaki stosowany był dla pojedynczych artykułów.

Prace badawcze ograniczone zostały do tekstów naukowych. Takie dokumenty, ze względu na wymogi formalne dotyczące ich tworzenia, cechują się wysoką precyzją i jednoznacznością stosowanych w nich terminów i pojęć, dzięki czemu ułatwione jest wskazanie prawidłowych słów kluczowych reprezentujących treść takiego tekstu.

Ze względu na konieczność przeprowadzenia wiarygodnej weryfikacji wyników automatycznej analizy i wyznaczenia słów kluczowych, do opracowania wybrane zostały te dokumenty, które były zaopatrzone przez autorów w słowa kluczowe.

3.4.1. Korpus tekstów

W celu przeprowadzenia badań został zgromadzony zestaw dokumentów, których treść złożyła się na badawczy korpus tekstów. Na potrzeby niniejszej książki przeprowadzony został szereg analiz frekwencyjnych na utworzonym zbiorze. Z powodu szczególnego uwzględnienia w pracy słów kluczowych, do bezpośredniej analizy wybrano tylko artykuły wyposażone przez autorów w zestawy słów charakteryzujących ich treść. Pozostałe artykuły, które nie posiadały autorskich słów kluczowych, zostały poddane analizie automatycznej, której wyniki zasiliły zbiór danych frekwencyjnych o tekstach z danego zakresu tematycznego.

Z punktu widzenia badań przeprowadzonych na potrzeby niniejszej książki zaletą artykułów jest stosunkowo nieduża długość ich tekstów, po-

zwalająca w stosunkowo krótkim czasie dokonać analizy i porównania większej liczby tekstów, niż miałoby to miejsce w przypadku książek. Zaletą krótkich form piśmienniczych jest również wysoka kondensacja treści i pominięcie wyrazów o niskiej wartości informacyjnej, które mogą występować w treści dłuższych form piśmienniczych.

Na zbiór analizowanych tekstów składają się prace z zakresu informacji naukowej i bibliologii oraz nauk ekonomicznych i zarządzania. Główny zrąb korpusu stanowią prace poświęcone nauce o informacji, natomiast artykuły związane z ekonomią i zarządzaniem, tworzące mniejszy zbiór, zostały wybrane w celu weryfikacji wyników badań przeprowadzonych na korpusie podstawowym.

Taki dobór materiałów został podyktowany względami praktycznymi. Artykuły naukowe cechuje zazwyczaj precyzyjny, jednoznaczny język[188], dzięki czemu ułatwione są operacje wstępnego przetwarzania tekstu do dalszych analiz. W badaniach nad automatycznym przetwarzaniem języka naturalnego poważny problem stanowi np. ustalenie prawidłowego znaczenia dla homonimów. W tekstach naukowych zjawisko homonimii jest zmarginalizowane, natomiast jeżeli spotyka się homonimy, należą one do klasy słownictwa najczęstszego, o małej wartości informacyjnej.

Zebrane dokumenty zostały poddane dwóm rodzajom analizy. Pierwszy rodzaj, kluczowy dla niniejszej książki, stanowiła analiza automatyczna. Proces automatycznego opracowania dokumentów został zaprezentowany w poprzednim rozdziale. Dodatkowo analizowane były słowa kluczowe powstałe w wyniku procesów kognitywnych. Były one wybierane zarówno przez autorów poszczególnych tekstów, jak i przez specjalistów w procesie opracowania rzeczowego. Możliwość porównania zestawów słów kluczowych wskazanych przez aplikację ze wskazanymi przez człowieka, pozwoliła wybrać metodę automatyczną, która generowała odpowiedzi jak najbardziej zbliżone do naturalnych.

3.4.2. Teksty z zakresu informacji naukowej i bibliologii

Na część korpusu związaną tematycznie z informacją naukową i bibliologią składają się publikacje z trzech roczników „Przeglądu Bibliotecznego" oraz trzech roczników „Zagadnień Informacji Naukowej" – dla obu tytułów są to lata 2005-2007. Oprócz tego przeanalizowane zostały tek-

[188] Chociażby ze względu na stosunkowo niewielki udział homonimów czy metafor w treści.

sty z publikacji konferencyjnych. Z obu wiodących czasopism analizowano wyłącznie artykuły, pomijając inne, zamieszczone w nich formy piśmiennicze; natomiast w materiałach konferencyjnych publikowane były wyłącznie artykuły. Teksty materiałów konferencyjnych uwzględnione w niniejszej książce pochodzą z okresu 2005-2008[189]. Wybrano wydawnictwa, które są dostępne w postaci elektronicznej:

1. *Biblioteki naukowe w kulturze i cywilizacji. Działania i codzienność. Materiały konferencyjne, Poznań 15-17 czerwca 2005*, pod red. H. Ganińskiej, t.1. Poznań: Biblioteka Główna Politechniki Poznańskiej 2005.
2. *Biblioteki naukowe w kulturze i cywilizacji. Działania i codzienność. Materiały konferencyjne, Poznań 15-17 czerwca 2005*, pod red. H. Ganińskiej, t.2. Poznań: Biblioteka Główna Politechniki Poznańskiej 2005.
3. *Biblioteki XXI wieku. Czy przetrwamy?: II Konferencja Biblioteki Politechniki Łódzkiej, Łódź, 19-21 czerwca 2006 r. Materiały konferencyjne*. Łódź: Politechnika Łódzka 2006.
4. *Informacja dla nauki a świat zasobów cyfrowych*, pod red. H. Ganińskiej. Poznań: Biblioteka Główna Politechniki Poznańskiej 2008.

Korpus tekstów fachowych utworzony na potrzeby niniejszej książki zawiera po wstępnym opracowaniu 848.650 tokenów pochodzących ze 183 artykułów. Po sprowadzeniu wyrazów do postaci podstawowej korpus liczy 38.272 leksemów o średniej częstości ok. 22 wystąpień. Frekwencją poniżej średniej częstości wystąpień cechuje się 34.695 leksemów, co stanowi 91% całego zbioru. W całym korpusie 53 leksemy (czyli 0,14% zbioru) ma częstości wystąpień powyżej 1000, z kolei 1049 wyrazów (2,74% wszystkich leksemów) występują częściej niż 100 razy, zaś 18.459 leksemów, co stanowi prawie połowę całego korpusu (dokładniej 48,23%), wystąpiło tylko jeden raz. Rozkład frekwencyjny pięćdziesięciu trzech leksemów o częstościach występowania powyżej 1000 prezentuje tabela 6.

[189] Podobny zakres chronologiczny badanego słownictwa zapewnia, że ogólne charakterystyki frekwencyjne dla tekstów będą wspólne, co z kolei umożliwi przeprowadzenie wiarygodnych i miarodajnych porównań kwantytatywnych.

Tabela 6. Częstości wystąpień oraz udział procentowy leksemów o frekwencjach powyżej 1000 w połączonym korpusie z zakresu informacji naukowej.

Lp.	Leksem	Liczba wystąpień	Udział procentowy w korpusie
1	biblioteka	11855	1,3969
2	informacja	5597	0,6595
3	dane	3507	0,4132
4	naukowy	3417	0,4026
5	praca	3119	0,3675
6	system	2394	0,2821
7	bibliotekarz	2012	0,2371
8	biblioteczny	1985	0,2339
9	użytkownik	1974	0,2326
10	elektroniczny	1945	0,2292
11	wiedza	1921	0,2264
12	polski	1871	0,2205
13	dokument	1722	0,2029
14	język	1660	0,1956
15	baza	1658	0,1954
16	zbiór	1609	0,1896
17	badać	1587	0,1870
18	czasopismo	1540	0,1815
19	dostęp	1535	0,1809
20	tworzyć	1483	0,1747
21	zasób	1408	0,1659
22	dostępny	1402	0,1652
23	informacyjny	1398	0,1647
24	cyfrowy	1370	0,1614
25	organizacja	1288	0,1518
26	artykuł	1284	0,1513

Lp.	Leksem	Liczba wystąpień	Udział procentowy w korpusie
27	bibliograficzny	1265	0,1491
28	Internet	1256	0,1480
29	publikacja	1236	0,1456
30	nazywać	1229	0,1448
31	proces	1222	0,1440
32	pracownik	1220	0,1438
33	katalog	1207	0,1422
34	formacja	1192	0,1405
35	bibliografia	1174	0,1383
36	nauka	1166	0,1374
37	stać	1156	0,1362
38	korzystać	1147	0,1352
39	cel	1134	0,1336
40	książek	1109	0,1307
41	osoba	1094	0,1289
42	narodowy	1084	0,1277
43	źródło	1081	0,1274
44	prowadzić	1074	0,1266
45	online	1065	0,1255
46	nowa	1054	0,1242
47	strona	1052	0,1240
48	forma	1052	0,1240
49	liczba	1051	0,1238
50	zakres	1040	0,1225
51	opis	1034	0,1218
52	stosować	1033	0,1217
53	działać	1008	0,1188

Źródło: Opracowanie własne na podstawie wyników działania aplikacji autorskiej.

Analizując skład zbioru leksemów o najwyższych frekwencjach można zauważyć liczny, zdecydowanie wyższy niż w korpusie ogólnym, udział jednostek leksykalnych związanych tematycznie z nauką o informacji. Spośród 53 leksemów występujących najczęściej w badanych tekstach 24 stanowią jednostki leksykalne związane z nią tematycznie. Rozkład częstości najpopularniejszych leksemów w ogólnym korpusie języka polskiego wśród 51 najczęściej występujących słowoform notuje wyłącznie słownictwo gramatyczne, z dwoma wyjątkami, którymi są leksemy **człowiek** oraz **praca**[190].

Podczas budowy korpusu badawczego z wymienionych dokumentów źródłowych dokonano analiz frekwencyjnych zarówno poszczególnych części składowych, jak i całego uzyskanego korpusu. W następnych podrozdziałach zostaną zaprezentowane charakterystyki frekwencyjne słownictwa poszczególnych części korpusu podstawowego.

3.4.2.1. Artykuły z czasopism

„Przegląd Biblioteczny" reprezentowany jest przez 50 artykułów liczących łącznie, po usunięciu słów nieznaczących, 322.104 wyrazy. Po sprowadzeniu wszystkich uwzględnionych słów z artykułów „PB" do postaci kanonicznej, uzyskano 19.577 leksemów. Średnia liczba wystąpień pojedynczego leksemu wynosi 16,4. W tabeli 7. przedstawiony został rozkład częstości trzydziestu słów występujących najczęściej w analizowanych artykułach.

Tabela 7. Frekwencje 30 najpopularniejszych słów pojawiających się w zoptymalizowanych lingwistycznie artykułach z „Przeglądu Bibliotecznego".

Lp.	Leksem	Liczba wystąpień	Udział procentowy w korpusie tekstów „PB"
1	biblioteka	4347	1,350
2	praca	1428	0,443
3	informacja	1327	0,412
4	naukowy	1296	0,402
5	dane	1055	0,328
6	polski	1046	0,325

[190] Por. *Lista słów*, dz. cyt.

Lp.	Leksem	Liczba wystąpień	Udział procentowy w korpusie tekstów „PB”
7	biblioteczny	1010	0,314
8	artykuł	955	0,296
9	książka	820	0,255
10	bibliotekarz	801	0,249
11	język	739	0,229
12	bibliografia	706	0,219
13	narodowy	696	0,216
14	użytkownik	693	0,215
15	badać	680	0,211
16	system	678	0,210
17	bibliograficzny	675	0,210
18	publikacja	648	0,201
19	organizacja	635	0,197
20	elektroniczny	624	0,194
21	dostęp	616	0,191
22	zbiór	582	0,181
23	Warszawa	576	0,179
24	instytut	565	0,175
25	działać	547	0,170
26	cel	545	0,169
27	dokument	545	0,169
28	tworzyć	532	0,165
29	zakres	517	0,161
30	przegląd	511	0,159

Źródło: Opracowanie własne na podstawie analizy artykułów „PB”.

Najczęściej występującym w analizowanych artykułach słowem jest leksem **biblioteka**, którego różne formy gramatyczne pojawiały się łącznie 4347 razy, co stanowi ok. 1,35% całego słownictwa tej części korpusu. Jest

to wartość o rząd wielkości przewyższająca kolejne, często występujące leksemy. Drugim, najczęściej występującym w tej części korpusu badawczego leksemem jest wyraz **praca**, pojawiający się łącznie 1428 razy, czyli stanowiący 0,44% słownictwa w zoptymalizowanych lingwistycznie artykułach „Przeglądu Bibliotecznego". Na liście frekwencyjnej *Korpusu języka polskiego PWN* leksem ten umiejscowiony jest na 44. pozycji, poprzedzają go głównie wyrazy ze strefy słownictwa gramatycznego oraz następujące wyrazy nienależące do tej strefy: **mieć** (poz. 16), **rok** (poz. 21), **wiedzieć** (poz. 36), **czas** (poz. 41) i **człowiek** (poz. 42)[191]. Wśród 30 słów najczęściej występujących w artykułach „PB" 18 leksemów, czyli 60% można zaliczyć do strefy słownictwa charakterystycznego, ściśle związanego znaczeniowo z zakresem tematycznym czasopisma i generalnie korpusu badawczego. Z kolei wśród 100 najpopularniejszych w tej części korpusu leksemów słownictwo charakterystyczne stanowi już tylko ok. 40%[192].

Kolejnym czasopismem branżowym, z którego zaczerpnięto artykuły do budowy korpusu, są „Zagadnienia Informacji Naukowej". Z roczników 2005-2007 pochodzi 39 artykułów o łącznej objętości 184.150 wyrazów po usunięciu słów mało znaczących. Po sprowadzeniu uzyskanych w procesie optymalizacji lingwistycznej wyrazów do postaci podstawowej zostało 15.949 leksemów o średniej liczbie wystąpień równej 11,55. Rozkład częstości wystąpień 30 najpopularniejszych leksemów dla tego czasopisma prezentuje tabela 8.

Tabela 8. Częstości wystąpień oraz udział procentowy trzydziestu najpopularniejszych leksemów w przetworzonym korpusie artykułów „Zagadnień Informacji Naukowej".

Lp.	Leksem	Liczba wystąpień	Udział procentowy w korpusie tekstów „ZIN"
1	biblioteka	1446	0,785
2	dane	1033	0,561
3	informacja	1012	0,550
4	system	832	0,452
5	język	817	0,444
6	wiedza	779	0,423

[191] Dane za: *Korpus języka polskiego Wydawnictwa...*, dz. cyt.

[192] Opracowanie własne na podstawie analizy artykułów wchodzących w skład korpusu badawczego.

Lp.	Leksem	Liczba wystąpień	Udział procentowy w korpusie tekstów „ZIN”
7	książka	775	0,421
8	dokument	593	0,322
9	naukowy	584	0,317
10	użytkownik	479	0,260
11	baza	473	0,257
12	opis	470	0,255
13	praca	458	0,249
14	dostęp	400	0,217
15	bibliograficzny	375	0,204
16	termin	375	0,204
17	tworzyć	372	0,202
18	zbiór	367	0,199
19	bibliografia	365	0,198
20	badać	364	0,198
21	nauka	351	0,191
22	liczba	347	0,188
23	czasopismo	339	0,184
24	dziać	319	0,173
25	osoba	305	0,166
26	źródło	298	0,162
27	polski	296	0,161
28	stosować	293	0,159
29	serwis	291	0,158
30	zasób	287	0,156

Źródło: Opracowanie własne na podstawie analizy artykułów „ZIN” przez aplikację autorską.

W przypadku artykułów „ZIN” również leksem **biblioteka** występował najczęściej, w tym fragmencie korpusu liczba jego wystąpień wynosi 1446, co stanowi ok. 0,8% słownictwa całej części. W przypadku tej

części korpusu badawczego różnica pomiędzy najczęściej występującym leksemem a pozostałymi leksemami na liście nie jest tak znaczna, jak w przypadku słownictwa z artykułów z „Przeglądu Bibliotecznego". Dla drugiego najpopularniejszego leksemu wśród słownictwa „Zagadnień Informacji Naukowej", **dane**, liczba wystąpień wynosi 1033, co stanowi ok. 0,56% całego słownictwa tej części korpusu. Do strefy słownictwa charakterystycznego wśród 30 najczęściej występujących leksemów omawianej części korpusu należy 18 leksemów (60%), czyli wartości równe odpowiedniemu fragmentowi korpusu artykułów „PB". Zaś wśród 100 najczęściej występujących leksemów „ZIN" ok. 38% stanowi słownictwo charakterystyczne[193].

Po wstępnym opracowaniu teksty z obu periodyków zostały połączone w jeden zbiór reprezentujący 89 artykułów o liczebności 506.254 wyrazów, które po sprowadzeniu do postaci podstawowej pozwoliły wygenerować 27.610 leksemów. Średnia częstość wystąpień jednego leksemu wynosi 18. Wielkości leksykalne obu zbiorów tekstów z czasopism są porównywalne, różnica wynosi 3628 leksemów, co również potwierdza[194], że słownictwo w obu periodykach stanowi jednolity zbiór[195]. Tabela 9. prezentuje frekwencje oraz udział procentowy w połączonych tekstach z czasopism dla trzydziestu najczęściej występujących leksemów:

Tabela 9. Częstości wystąpień oraz udział procentowy trzydziestu najpopularniejszych leksemów w połączonych tekstach artykułów „Zagadnień Informacji Naukowej" i „Przeglądu Bibliotecznego".

Lp.	Leksem	Liczba wystąpień	Udział procentowy w korpusie połączonych tekstów „PB" i „ZIN"
1	biblioteka	5793	1,144
2	informacja	2339	0,462
3	dane	2088	0,412
4	praca	1886	0,373
5	naukowy	1880	0,371

[193] Opracowanie własne na podstawie analizy artykułów wchodzących w skład korpusu badawczego tekstów.

[194] Innym pośrednim dowodem podobieństwa słownictwa jest średni udział procentowy, który dla obu tytułów jest bardzo zbliżony.

[195] Opracowanie własne na podstawie analizy artykułów wchodzących w skład korpusu badawczego.

Lp.	Leksem	Liczba wystąpień	Udział procentowy w korpusie połączonych tekstów „PB" i „ZIN"
6	książka	1595	0,315
7	język	1556	0,307
8	system	1510	0,298
9	polski	1342	0,265
10	wiedza	1251	0,247
11	biblioteczny	1181	0,233
12	użytkownik	1172	0,232
13	dokument	1138	0,225
14	artykuł	1130	0,223
15	bibliografia	1071	0,212
16	bibliograficzny	1050	0,207
17	badać	1044	0,206
18	dostęp	1016	0,201
19	bibliotekarz	957	0,189
20	publikacja	934	0,184
21	tworzyć	904	0,179
22	elektroniczny	875	0,173
23	narodowy	873	0,172
24	organizacja	840	0,166
25	warszawa	837	0,165
26	baza	818	0,162
27	opis	808	0,160
28	stosować	801	0,158
29	zakres	792	0,156
30	osoba	788	0,156

Źródło: Opracowanie własne na podstawie analizy artykułów „PB" i „ZIN" przez aplikację autorską.

3.4.2.2. Artykuły z materiałów konferencyjnych

Jak już wspomniano wcześniej, w skład korpusu badawczego związanego tematycznie z zagadnieniami informacji naukowej i bibliologii włączono również artykuły z prac zbiorowych i publikacji pokonferencyjnych. Na tę część korpusu składają się teksty 94 artykułów o łącznej objętości 342.396 słów. Po sprowadzeniu wszystkich jednostek leksykalnych do postaci kanonicznej uzyskano 21.769 leksemów. Średnia liczba wystąpień wynosi ok. 16 (dokładnie 15,73). Z publikacji owych pochodziły również pojedyncze artykuły, które zostały poddane szczegółowej analizie podczas trzeciej, automatycznej fazy procesu badawczego opisywanego w niniejszej książce. Wyniki analiz dla artykułów zostaną opisane w dalszej części rozdziału, natomiast w tym miejscu zaprezentowane zostaną charakterystyki artykułowej części korpusu oraz charakterystyki całego korpusu tekstów z zakresu informacji naukowej i bibliotekoznawstwa.

Tabela 10. prezentuje dane dotyczące frekwencji 30 najpowszechniej występujących słów w korpusie artykułów.

Tabela 10. Częstości wystąpień oraz udział procentowy trzydziestu najpopularniejszych leksemów w zbiorze tekstów artykułów z materiałów konferencyjnych.

Lp.	Leksem	Liczba wystąpień	Udział procentowy w zbiorze tekstów z artykułów konferencyjnych
1	biblioteka	6062	1,770
2	użytkownik	1600	0,467
3	naukowy	1537	0,449
4	dane	1419	0,414
5	praca	1233	0,360
6	informacja	1149	0,336
7	Internet	1140	0,333
8	elektroniczny	1070	0,313
9	bibliotekarz	1055	0,308
10	formacja	919	0,268
11	system	884	0,258

Lp.	Leksem	Liczba wystąpień	Udział procentowy w zbiorze tekstów z artykułów konferencyjnych
12	cyfrowy	849	0,248
13	baza	840	0,245
14	biblioteczny	804	0,235
15	czasopismo	798	0,233
16	dostępny	750	0,219
17	informacyjny	718	0,210
18	zasób	712	0,208
19	wiedza	670	0,196
20	katalog	665	0,194
21	zbiór	660	0,193
22	pracownik	638	0,186
23	dokument	584	0,171
24	http	584	0,171
25	tworzyć	579	0,169
26	badać	543	0,159
27	korzystać	536	0,157
28	polski	529	0,154
29	uczelnia	527	0,154
30	online	526	0,154

Źródło: Opracowanie własne na podstawie analizy artykułów materiałów konferencyjnych przez aplikację autorską.

Również w przypadku artykułów konferencyjnych najczęściej występującym słowem jest **biblioteka**, która pojawia się w tym zbiorze tekstów 6062 razy, co daje udział procentowy na poziomie 1,77%. Także w tym zestawie tekstów leksem **biblioteka** przewyższa o rząd wielkości frekwencję kolejnych bardzo częstych leksemów. Drugim, najczęściej występującym w tej części korpusu badawczego słowem jest wyraz **użytkownik**, pojawiający się łącznie 1600 razy, co stanowi 0,47% słownictwa w zoptymalizowanych lingwistycznie artykułach z materiałów konferencyjnych.

3.4.3. Subkorpus ekonomia i zarządzanie

Oprócz tekstów poświęconych informacji naukowej badaniu poddano również artykuły z zakresu nauk ekonomicznych i zarządzania. Ten zbiór został utworzony głównie w celu weryfikacji ustaleń analiz przeprowadzonych na tekstach z zakresu informacji naukowej i bibliologii. Teksty te stanowiły mniejszy zestaw, łączna objętość wynosi ok. 195.300 tokenów pochodzących z 54 artykułów.

Artykuły, których treść utworzyła zbiór tekstów z zakresu ekonomii i zarządzania, opublikowane zostały w następujących wydawnictwach:

1. *Zarządzanie XXXVII. Nauki humanistyczno-społeczne*, z. 387, pod red. M. Haffera. Toruń: Wydawnictwo UMK 2007.
2. *Ekonomia XXXVIII. Nauki humanistyczno-społeczne*, z. 388, pod red. M. Piłatowskiej. Toruń: Wydawnictwo UMK 2008.
3. *Ekonomia XXXIX. Nauki humanistyczno-społeczne, zeszyt specjalny. Dynamiczne modele ekonometryczne*, z. 389, pod red. M. Piłatowskiej. Toruń: Wydawnictwo UMK 2009.

Po sprowadzeniu wyrazów do postaci podstawowej zbiór liczy 10.228 leksemów o średniej częstości występowania ok. 19 (19,09). Wśród analizowanych tekstów z zakresu ekonomii i zarządzania 8978 leksemów, co stanowi 88% całego słownika, cechuje się częstością wystąpień poniżej średniej. W tym zbiorze jedynie leksem **model** ma frekwencję wyższą niż 1000 wystąpień, zaś 195 leksemów (co stanowi 1,91% zbioru słownictwa) występuje częściej niż 100 razy.

W obu badanych zbiorach leksemy o wysokich częstościach występowania stanowią podobną procentowo część subkorpusu, odpowiednio 3% i 2% słowników. Podobnie kształtują się również udziały słownictwa rzadkiego, o częstościach występowania poniżej średniej, odpowiednio 91% dla korpusu tekstów informacji naukowej i 88% dla tekstów z zakresu zarządzania. Jednakże istotna różnica występuje w podzbiorach leksemów pojawiających się tylko raz. W obu zbiorach tekstów leksemy takie stanowią spore części, jednakże w tekstach z zakresu informacji naukowej ich udział wynosi aż 48%, zaś dla tekstów ekonomicznych ok. 37% (dokładnie 37,39%).

Ta różnica może wskazywać na bogatszy zasób słownictwa występujący w tekstach związanych z nauką o informacji oraz pośrednio intuicyjne założenia, że teksty humanistyczne są bardziej zróżnicowane treściowo niż teksty z zakresu nauk ścisłych.

W związku z brakiem autorskich słów kluczowych przy większości analizowanych tekstów, nie wszystkie dokumenty z zakresu informacji naukowej i bibliologii zostały analizowane w tym i w kolejnym etapie doświadczenia. Jednakże, nawet pomimo niewykorzystania bezpośredniego owych dokumentów do analizy, stanowiły one doskonały materiał do zbudowania korpusu porównawczego. Dzięki ograniczeniu zakresu tematycznego zawartych artykułów do zagadnień jednej dziedziny, dostarczały cennych informacji o rozkładzie częstości wystąpień poszczególnych słów, pozwalając tym samym zbudować wartościową i wiarygodną podstawę do dalszych porównań i oszacowań wagi słów. Informacje uzyskane z analizy tych dokumentów wykorzystywane były podczas następnej fazy prac badawczych.

3.4.4. Słowa kluczowe

Zgodnie z definicją słów kluczowych jako wyrażeń charakteryzujących treść dokumentu, powinny być to wyrazy o dużej wartości informacyjnej. Uwzględniając prawa statystyczne dotyczące języków naturalnych, słów kluczowych należy szukać wśród słownictwa strefy charakterystycznej oraz częściowo wśród słownictwa rzadkiego. Z pewnością na słowa reprezentujące treść dokumentu nie kwalifikują się słowa ze strefy słownictwa częstego oraz ze strefy gramatycznej. Słowa częste są zbyt powszechnie wykorzystywane, występują w tekstach o różnorakiej tematyce, żeby w jak najbardziej jednoznaczny sposób reprezentowały treść danego dokumentu[196]. Podobnie słowa pełniące funkcję gramatyczną w wypowiedziach języka naturalnego są silnie nacechowane informacją, ale jest to informacja pozatreściowa, przez co wyrazy takie również niejednoznacznie określałyby treść dokumentu.

3.4.4.1. Słowa kluczowe wybierane przez autorów

Słowa kluczowe ustalone przez autorów, pobrane z analizowanych tekstów, zostały zapisane w sposób umożliwiający identyfikację dokumentu, z którego pochodzą. Z powodu braku tego typu słów w niektórych dokumentach, szczególnie z grupy głównej, nie wszystkie teksty włączone w skład korpusu podlegały opracowaniu na tym etapie. Przyczyny takiej

[196] Por. I. Kamińska-Szmaj: dz. cyt., s. 23.

sytuacji zostały opisane w dalszej części rozdziału, w sekcji poświęconej problemom napotkanym podczas selekcji dokumentów.

Ponieważ publikacje naukowe przygotowywane są zgodnie z ustalonymi, uniwersalnymi zasadami, można założyć, że słowa kluczowe wskazane przez autorów rzeczywiście prawidłowo reprezentują treść artykułów. Należy jednakże mieć na uwadze również specyfikę tworzenia autorskich zestawów słów kluczowych. Każda publikacja naukowa przygotowywana jest zgodnie z profilem wydawnictwa, do którego zostanie zgłoszona lub przez które została zamówiona. Skutkuje to dodaniem słów kluczowych uwypuklających aspekty treści artykułu związane z profilem docelowego wydawnictwa lub odbiorcy. Jest to praktyka zgodna z zaleceniami opracowania rzeczowego dokumentów. W opracowaniu tego typu wskazuje się zarówno przedmiot dokumentu, jak i aspekt zaprezentowania tego przedmiotu w danej pozycji piśmienniczej[197]. Odpowiednia analiza treści dokumentu oraz rozkładu częstości występujących leksemów może wskazać jednostki leksykalne, które są dobrymi reprezentacjami całej treści dokumentu. Przykładowe zestawy słów kluczowych autorskich przypisanych do artykułów z głównej części materiału badawczego zaprezentowane zostały w tabeli 11.

Tabela 11. Wybrane zestawy autorskich słów kluczowych z głównego zrębu materiału badawczego.

Identyfikator dokumentu	Tytuł	Autorskie słowa kluczowe
1a	*BazTOL jako przykład serwisu typu* subject gateway *o kontrolowanej jakości*	baza danych; nauki techniczne; portal; serwis tematyczny o kontrolowanej jakości; wyszukiwanie informacji w Internecie
1b	*Biblioteka akademicka jako element globalnej cyfrowej infrastruktury informacyjnej*	biblioteki akademickie; globalna infrastruktura informacyjna

[197] Por. T. Głowacka: *Kartoteka wzorcowa języka KABA. Stosowanie w katalogowaniu przedmiotowym*. Warszawa: SBP 1997.

Identyfikator dokumentu	Tytuł	Autorskie słowa kluczowe
1d	*Biblioteka Jagiellońska 21. wieku – tradycja i nowoczesność*	Biblioteka Jagiellońska; budownictwo biblioteczne; komputeryzacja bibliotek naukowych; NUKAT; współpraca bibliotek; zbiory biblioteczne - ochrona
1f	*Bibliotekarze dyplomowani – wczoraj, dziś i jutro*	bibliotekarstwo - Polska; bibliotekarze - status prawny
1g	*Biblioteki akademickie – trendy dotyczące zasobów elektronicznych*	archiwizacja Web; archiwizacja; czasopisma wolnodostępne; digitalizacja; kolekcje elektroniczne; konserwacja; repozytorium instytucjonalne; usługi informacyjne
1h	*Biblioteki naukowe wobec kulturowych i cywilizacyjnych potrzeb społeczeństwa*	biblioteka hybrydowa; biblioteka naukowa; centrum edukacji, informacji i kultury; Internet
1k	*Elektroniczna książka na elektronicznym papierze. Czy to już zmierzch ery druku?*	badanie użytkowników; e-czytnik; e-ink; elektroniczny papier; e-papier
1p	*Od witryny internetowej do portalu bibliotecznego*	metadane; portal biblioteczny; technologia informacji; zasoby internetowe
1s	*Otwarte zasoby wiedzy na stronach internetowych wybranych bibliotek uczelni technicznych w Polsce i na świecie – przegląd z perspektywy doświadczeń Wypożyczalni Międzybibliotecznej Biblioteki Głównej Politechniki Warszawskiej*	wypożyczanie międzybiblioteczne; Biblioteka Główna Politechniki Warszawskiej; Internet; otwarte zasoby wiedzy;
1u	*Podlaska Biblioteka Cyfrowa*	biblioteka cyfrowa; digitalizacja

Źródło: Opracowanie własne na podstawie wyników automatycznej analizy dokumentów poddanych badaniom.

Dodatkowym elementem badania było sprawdzenie, w jaki sposób osoby profesjonalnie zaznajomione z zasadami indeksowania treści dokumentów opiszą za pomocą słów kluczowych treść tekstów analizowanych na potrzeby niniejszej książki.

3.4.4.2. Słowa kluczowe wskazane przez indeksatorów

Kolejny etap procedury badawczej obejmował analizę treści artykułów wykonaną przez człowieka. Nieocenioną pomoc stanowili tu studenci Instytutu Informacji Naukowej i Bibliologii oraz studenci Instytutu Archiwistyki UMK w Toruniu. Osoby biorące udział w tej fazie badań wykazywały się już pewnym doświadczeniem praktycznym w opracowaniu zarówno formalnym, jak i rzeczowym dokumentów. Doświadczenie to zdobyte było m.in. podczas zajęć *Rzeczowe opracowanie dokumentów – UKD*, *Rzeczowe opracowanie dokumentów – JHP KABA*, czy *Opracowanie formalne dokumentów*. Dodatkowo przydatne okazały się na tym etapie procesu badawczego doświadczenia wyniesione przez uczestników z obowiązkowych praktyk w bibliotekach (w tym naukowych), archiwach bądź ośrodkach informacji. Dzięki temu osoby biorące udział w procesie opracowania rzeczowego dokumentów z dużą dozą wiarygodności ustalały słowa kluczowe reprezentujące treść analizowanych przez siebie dokumentów. W grupie uczestników liczącej 60 osób 64% stanowiły kobiety a 36% – mężczyźni. Średnia wieku członków grupy wynosiła około 20 lat.

Aby zachować jak największy obiektywizm w ocenie ustalonych słów kluczowych, każdy artykuł był opisywany niezależnie przez kilka osób. Wyniki ostateczne tego typu analizy zostały połączone w jeden opis, z usunięciem powtarzających się terminów w taki sposób, aby każde użyte wyrażenie występowało w opisie tylko raz. W przypadkach wątpliwych, gdy zaproponowane słowa kluczowe odbiegały w istotny sposób od pozostałych propozycji, były pomijane w zestawieniu zbiorczym.

Ten etap badań przeprowadzony został w listopadzie 2010 roku. W jego wyniku, po zweryfikowaniu odpowiedzi uczestników, udało się uzyskać 40 zestawów wyrażeń charakteryzujących treść przypisanych średnio do 20 różnych artykułów, czyli 800 niezależnych zestawów słów kluczowych. Po opracowaniu uzyskanych w ten sposób słów kluczowych ostatecznie pozostało 200 opisów. Przykładowe zestawy słów kluczowych zaproponowanych w procesie indeksacji przez specjalistów zaprezentowane zostały w tabeli 12.

Tabela 12. Przykładowe słowa kluczowe uzyskane w tradycyjnym procesie opracowania rzeczowego treści dokumentów.

Identyfikator dokumentu	Tytuł	Słowa kluczowe wskazane przez indeksatorów
1a	*BazTOL jako przykład serwisu typu* subject gateway *o kontrolowanej jakości*	baza danych; nauki techniczne; katalogowanie zasobów sieciowych; portal; serwis tematyczny o kontrolowanej jakości; struktura serwisu; wyszukiwanie informacji w internecie
1b	*Biblioteka akademicka jako element globalnej cyfrowej infrastruktury informacyjnej*	biblioteki akademickie; biblioteki cyfrowe; cyfrowa infrastruktura informacyjna; digitalizacja; formy kształcenia; globalizacja; globalna infrastruktura informacyjna; Internet; katalogowanie; kształcenie bibliotekarzy; nauczanie zdalne; organizacja przestrzeni bibliotecznej; rola bibliotek w nauczaniu ustawicznym
1d	*Biblioteka Jagiellońska 21. wieku – tradycja i nowoczesność*	biblioteka jagiellońska; digitalizacja zbiorów; informacja naukowa; komputeryzacja bibliotek; ochrona zbiorów
1f	*Bibliotekarze dyplomowani – wczoraj, dziś i jutro*	bibliotekarze w Polsce; bibliotekarze dyplomowani
1g	*Biblioteki akademickie – trendy dotyczące zasobów elektronicznych*	archiwizacja informacji; biblioteki akademickie; biblioteki cyfrowe; digitalizacja; e-biblioteki; e-czasopisma; e-kolekcje; elektroniczne repozytoria; e-zbiory; konserwacja zbiorów; naukowe źródła informacji; technologie cyfrowe; technologie informacyjne; usługi informacyjne; zasoby elektroniczne; zasoby informacyjne; źródła informacji

Identyfikator dokumentu	Tytuł	Słowa kluczowe wskazane przez indeksatorów
1h	*Biblioteki naukowe wobec kulturowych i cywilizacyjnych potrzeb społeczeństwa*	biblioteka hybrydowa; biblioteka naukowa; dziedzictwo kulturowe; edukacja; Internet; kultura; nauka; społeczeństwo informacyjne; technologie informacyjno-komunikacyjne
1k	*Elektroniczna książka na elektronicznym papierze. Czy to już zmierzch ery druku?*	badania ankietowe użytkowników; e-czytnik; elektroniczna książka; e-papier; książka elektroniczna; papier elektroniczny
1p	*Od witryny internetowej do portalu bibliotecznego*	biblioteczne portale internetowe; Biblioteka Główna Politechniki Poznańskiej; dostęp do zasobów informacji; informacja naukowa; Internet; metadane; portal biblioteczny; rywalizacja dostawców informacji; technologia informacyjna; witryna internetowa; zadania portalu bibliotecznego; zasoby internetowe; zasób elektroniczny
1s	*Otwarte zasoby wiedzy na stronach internetowych wybranych bibliotek uczelni technicznych w Polsce i na świecie – przegląd z perspektywy doświadczeń Wypożyczalni Międzybibliotecznej Biblioteki Głównej Politechniki Warszawskiej*	Biblioteka Główna Politechniki Warszawskiej; Internet; metody wyszukiwania dokumentów; otwarte zasoby cyfrowe; otwarte zasoby wiedzy; procedury wyszukiwawcze; strony internetowe bibliotek; wypożyczanie międzybiblioteczne; zasoby naukowe
1u	*Podlaska Biblioteka Cyfrowa*	biblioteki cyfrowe; digitalizacja; Podlaska Biblioteka Cyfrowa; udostępnianie zbiorów; wdrażanie technologii informacyjnych; zadania biblioteki

Źródło: Opracowanie własne na podstawie odpowiedzi indeksatorów.

Kolejny etap badań dostarczał w sposób zautomatyzowany informacji frekwencyjnych na temat słownictwa wykorzystanego w analizowanych dokumentach.

3.4.4.3. Słowa kluczowe generowane automatycznie

Ostatnim etapem prac badawczych było przeprowadzenie analizy frekwencyjnej słownictwa z poszczególnych artykułów oraz automatyczne wskazanie na podstawie uzyskanych wyników leksemów – kandydatów na słowa kluczowe reprezentujące treść dokumentu. Wyselekcjonowane artykuły, gdzie warunkiem koniecznym było posiadanie autorskich słów kluczowych, zostały poddane procesowi wstępnego opracowania lingwistycznego. Po przeprowadzeniu operacji wstępnych słownictwo z poszczególnych artykułów było analizowane frekwencyjnie. Badane były zależności pomiędzy liczbą wystąpień w pojedynczym artykule oraz w całej ich kolekcji. Porównania te miały na celu ustalenie wagi poszczególnych słów w artykule i wskazanie słów kluczowych dla danego tekstu. Należy podkreślić pewną różnicę pomiędzy wskazywaniem słów kluczowych przez człowieka a przeprowadzeniem podobnego procesu przez aplikację. Większość słów kluczowych wskazywanych przez człowieka stanowiły wyrażenia kilkuwyrazowe, natomiast program komputerowy wskazywał zawsze słowa kluczowe jednowyrazowe. Istnieją co prawda sposoby wskazywania tzw. kolokacji, czyli powiązań wyrażeniowych między poszczególnymi elementami tekstu, ale wykraczają one poza zakres badawczy niniejszej książki.

Poszczególne etapy przygotowawcze tego procesu zostały opisane we wcześniejszych rozdziałach. Dla przypomnienia zamieszczono tu skrócony opis operacji. Wstępnie oczyszczony tekst (po usunięciu słów nieznaczących) poddawany jest operacji lematyzacji, w wyniku której powstaje słownik frekwencyjny leksemów dla indywidualnego artykułu. Słownik taki jest podstawą dalszych operacji analizy i porównań. Ze względu na wymogi niektórych zastosowanych wzorów, dla każdego słowa dodatkowo przechowywane były informacje o liczbie dokumentów, w których słowo wystąpiło. Z kolei porównanie częstości wystąpienia słowa do jego frekwencji sumarycznych wymagało dostępu do słownika frekwencyjnego lematów w skali całego korpusu.

Wyniki, w postaci listy potencjalnych słów kluczowych, uzyskane jako efekt zastosowania odpowiednich analiz frekwencyjnych porównywane były ze słowami kluczowymi wskazanymi w dwóch poprzednich etapach badań. Aplikacja stworzona na potrzeby niniejszej książki porządkowała wszystkie analizowane leksemy tworzące tekst poszczególnych dokumentów według ogólnej liczby wystąpień w konkretnym tekście. Z tego powodu wśród słowoform o najwyższych pozycjach rankingowych[198] można upatrywać leksemów – kandydatów na słowa kluczowe. Dzięki wykluczeniu z operacji automatycznego przetwarzania słów występujących powszechnie w języku polskim (cechujących się największymi częstościami w tekstach języka polskiego) pozostałe wyrazy o wysokich frekwencjach można zaliczyć do grona słów charakterystycznych dla danej kategorii słownictwa. Założenie to znajduje potwierdzenie w analizie list rankingowych zarówno dla tekstów z zakresu informacji naukowej, jak i dotyczących zagadnień ekonomii i zarządzania, co zostało zaprezentowane w poprzednich podrozdziałach.

Oczywiście samo ustalenie frekwencji i posortowanie leksemów ze względu na wartość tej cechy nie może i nie stanowi dostatecznego sposobu na wskazanie słów kluczowych. Niemniej jednak, liczba wystąpień wyrazu w danym dokumencie stanowiła podstawę do dalszych operacji arytmetyczno-statystycznych pozwalających ustalić wagę danego leksemu w analizowanym tekście. W tabeli 13. zaprezentowano wygenerowane automatycznie listy rankingowe dwudziestu najczęstszych lematów, kandydujących na słowa kluczowe dla przykładowych dokumentów z zakresu nauki o informacji. Proponowane słowa kluczowe zaprezentowane zostały w sposób zachowujący kolejność rankingową wyników działania aplikacji.

[198] Pozycja rankingowa danego leksemu zależy od częstości jego występowania w danym tekście – im częściej wyraz występował, tym wyższą pozycję zajmował na liście rankingowej. Tym samym wyraz najczęstszy w danym tekście miał pozycję 1., kolejne, rzadsze słowa zajmowały niższe pozycje rankingowe.

Tabela 13. Przykładowe słowa kluczowe uzyskane w procesie automatycznej analizy tekstów.

Identyfikator dokumentu	Tytuł	Słowa kluczowe generowane automatycznie
1a	*BazTOL jako przykład serwisu typu* subject gateway *o kontrolowanej jakości*	zasób; źródło; redaktor; baztol; katalogować; internet; baza; informacja; rekord; dziedzina; dostępny; portal; dostęp; online; serwis; wyszukiwać; tworzyć; politechnika; edycja
1b	*Biblioteka akademicka jako element globalnej cyfrowej infrastruktury informacyjnej*	biblioteka; zasób; cyfrowy; udostępniać; digitalizacja; infrastruktura; dokument; informacyjny; zbiór; oczywisty; globalny; naukowy; bibliotekarz; czytelnik; nośnik; dotyczyć; problem; zadać; proces; zmiana; bibliotekarz
1d	*Biblioteka Jagiellońska 21. wieku – tradycja i nowoczesność*	biblioteka; zbiór; katalog; naukowy; jagielloński; informacja; system; czasopismo; baza; czytelnik; polski; dokument; praca; elektroniczny; język; tworzyć; gmach; opracować; dostępny; komputeryzacja
1f	*Bibliotekarze dyplomowani – wczoraj, dziś i jutro*	bibliotekarz; dyplomowany; naukowy; egzamin; praca; pracownik; wyższy; biblioteka; ustawa; kandydat; grupa; biblioteczny; bibliotekarski; komisja; szkolnictwo; sprawa; język; stanowisko; prawo; status
1g	*Biblioteki akademickie – trendy dotyczące zasobów elektronicznych*	elektroniczny; biblioteka; informacja; czasopismo; cyfrowy; zbiór; dostęp; zasób; kolekcja; archiwizacja; usługa; naukowy; materiał; informacyjny; drukować; źródło; wolny; web; proces; open

Identyfikator dokumentu	Tytuł	Słowa kluczowe generowane automatycznie
1h	*Biblioteki naukowe wobec kulturowych i cywilizacyjnych potrzeb społeczeństwa*	biblioteka; informacja; użytkownik; cyfrowy; informacyjny; hybrydowy; *library*; technologia; dostęp; bibliotekarz; usługa; społeczeństwo; tradycyjny; wiedza; rozwój; zasób; przestrzeń; system; potrzeba; czas
1k	*Elektroniczna książka na elektronicznym papierze. Czy to już zmierzch ery druku?*	papier; czytnik; osoba; pytać; tabela; respondent; badać; książka; ankieta; elektroniczny; tekst; dostępny; wyświetlacz; grupa; strona; wzrok; czytelność; dokument; e-ink; odpowiedź
1p	*Od witryny internetowej do portalu bibliotecznego*	portal; biblioteczny; biblioteka; zasób; usługa; użytkownik; informacja; dostęp; online; katalog; baza; wyszukiwać; system; management; internetowy; dokument; zarządzać; standard; możliwość; integracja
1s	*Otwarte zasoby wiedzy na stronach internetowych wybranych bibliotek uczelni technicznych w Polsce i na świecie – przegląd z perspektywy doświadczeń Wypożyczalni Międzybibliotecznej Biblioteki Głównej Politechniki Warszawskiej*	biblioteka; zasób; otwarty; techniczny; uczelnia; zamówić; politechnika; czasopismo; naukowy; materiał; zbiór; cyfrowy; czytelnik; elektroniczny; katalog; strona; baza; dokument; dostęp; nauka
1u	*Podlaska Biblioteka Cyfrowa*	biblioteka; cyfrowy; podlaski; zbiór; digitalizacja; politechnika; liczba; naukowy; białostocki; dokument; konsorcjum; publikacja; Warszawa; kolekcja; dostępny; biblioteczny; rysunek; zasób; źródło; praca

Źródło: Opracowanie własne na podstawie wyników analiz automatycznych.

Analiza oraz interpretacja materiału badawczego i wyników badań

Jak już wspomniano wcześniej, główna część korpusu nadawczego tekstów została utworzona ze słownictwa występującego w artykułach z zakresu informacji naukowej i bibliologii. Artykuły te pochodziły z dwóch wiodących czasopism fachowych: „Przeglądu Bibliotecznego" oraz „Zagadnień Informacji Naukowej", a także z publikacji konferencyjnych. W bieżącym rozdziale zostanie zaprezentowana analiza i interpretacja wyników analiz, które zostały przeprowadzone na kolekcji zebranej tekstów.

4.1. Analiza głównego korpusu tekstów

Stworzony na potrzeby niniejszej książki korpus tekstów został poddany analizom, które opisano w poprzednim rozdziale. W kolejnych podrozdziałach przeprowadzono dyskusję oraz interpretację uzyskanych wyników.

4.1.1. Czasopisma

Operując na danych dla wydzielonych, niezależnych części tekstów z obu czasopism oraz dla połączonych tekstów można przeprowadzić analizę porównawczą słownictwa występującego w artykułach. Porównanie częstości dla dwudziestu najpowszechniejszych leksemów występujących w analizowanych artykułach z obu czasopism przedstawia tabela 14. Ponieważ obie analizowane części zbioru tekstów z zakresu nauki o informacji różniły się między sobą całkowitą liczebnością słów, w celach obiektyw-

nego porównania danych zarówno w tabelach, jak i w porównaniach opisowych, zaprezentowane są wartości udziałów procentowych poszczególnych leksemów w danej części korpusu badawczego. Wskazanie udziału procentowego w danym zbiorze pozwoli na obiektywne porównanie danych o różnych wielkościach bezwzględnych. W kolejnych porównaniach danych frekwencyjnych dotyczących analizowanych tekstów, miarą odniesienia będą właśnie udziały procentowe leksemów w zbiorach tekstów.

Tabela 14. Porównanie udziałów procentowych wystąpień dwudziestu najczęściej występujących leksemów w artykułach z „Przeglądu Bibliotecznego" i „Zagadnień Informacji Naukowej".

„Przegląd Biblioteczny"			„Zagadnienia Informacji Naukowej"		
Lp.	**Leksem**	**Udział procentowy**	**Lp.**	**Leksem**	**Udział procentowy**
1	**biblioteka**	1,3495	1	**biblioteka**	0,7852
2	**praca**	0,4433	2	**dane**	0,5609
3	**informacja**	0,4119	3	**informacja**	0,5495
4	**naukowy**	0,4023	4	**system**	0,4518
5	**dane**	0,3275	5	**język**	0,4436
6	polski	0,3247	6	wiedza	0,4230
7	biblioteczny	0,3135	7	**książka**	0,4208
8	artykuł	0,2964	8	dokument	0,3220
9	**książka**	0,2545	9	**naukowy**	0,3171
10	bibliotekarz	0,2486	10	**użytkownik**	0,2601
11	**język**	0,2294	11	baza	0,2568
12	**bibliografia**	0,2191	12	opis	0,2552
13	narodowy	0,2160	13	**praca**	0,2487
14	**użytkownik**	0,2151	14	dostęp	0,2172
15	**badać**	0,2111	15	**bibliograficzny**	0,2037
16	**system**	0,2104	16	termin	0,2036
17	**bibliograficzny**	0,2095	17	tworzyć	0,2020
18	publikacja	0,2011	18	zbiór	0,1992

„Przegląd Biblioteczny"			„Zagadnienia Informacji Naukowej"		
Lp.	**Leksem**	**Udział procentowy**	**Lp.**	**Leksem**	**Udział procentowy**
19	organizacja	0,1971	19	**bibliografia**	0,1982
20	elektroniczny	0,1937	20	**badać**	0,1976

Źródło: Opracowanie własne na podstawie analizy artykułów z „PB" i „ZIN" przez aplikację autorską.

W tabeli 14. za pomocą wytłuszczenia wyróżniono leksemy zgodne, o takich samych reprezentacjach graficznych i takim samym znaczeniu, występujące w obu czasopismach. Wśród 20 leksemów o najwyższych frekwencjach dla obu czasopism 12 z nich, czyli 60%, pokrywa się wzajemnie. Są to następujące leksemy:

- biblioteka,
- praca,
- informacja,
- naukowy,
- dane,
- książka,
- język,
- bibliografia,
- użytkownik,
- badać,
- system,
- bibliograficzny.

Jednakże można zauważyć różnice w dystrybucji częstości poszczególnych leksemów w tekstach z obu periodyków. Jedynie dwie słowoformy: **biblioteka** oraz **informacja** mają porównywalną wagę frekwencyjną (występują na tych samych pozycjach obu zestawień) wśród dwudziestu najpopularniejszych leksemów dla obu części[199]. Pozostałe jednostki leksykalne powtarzające się w tekstach z obu czasopism cechują się różnymi wagami w poszczególnych zbiorach.

Ogólnie, w traktowanych niezależnie zbiorach tekstów z obu czasopism 7826 leksemów występujących w tekstach „PB" pojawia się również w tek-

[199] Opracowanie własne na podstawie analizy artykułów wchodzących w skład korpusu badawczego.

stach „ZIN". Zatem słownictwo „Przeglądu Bibliotecznego" w 49% pokrywa leksykę występującą w dokumentach publikowanych w „Zagadnieniach Informacji Naukowej". Z drugiej zaś strony, 7716 leksemów z tekstów „ZIN" pojawia się w tekstach „PB", co daje 39% pokrycia leksyki. Średnio w treści obu czasopism pokrywa się ok. 7771 leksemów, czyli ok. 44% leksyki w dwóch analizowanych periodykach w latach 2005-2007 jest wspólne.

Dla 20 najczęstszych leksemów w obu fragmentach korpusu średnia liczba wystąpień oraz uśredniony procentowy udział wystąpień wynosi odpowiednio 1043 razy i 0,3237% dla „PB" oraz 619 razy i 0,3358% dla „ZIN". Zatem można stwierdzić, że średni udział procentowy najpopularniejszych leksemów w tekstach obu czasopism jest porównywalny, natomiast średnia liczba wystąpień wskazuje na wyższe zagęszczenie (wyższą stereotypowość) słownictwa strefy charakterystycznej w artykułach „Przeglądu Bibliotecznego". Dla słownictwa „PB" średnia liczba wystąpień wśród 20 leksemów o najwyższych frekwencjach jest prawie dwa razy większa (dokładnie 1,7) niż w odpowiedniej części słownictwa artykułów „ZIN", zaś średni udział procentowy słownictwa „Przeglądu" jest tylko o 0,0121% mniejszy niż w tekstach „Zagadnień". Z kolei wśród 100 słów o najwyższych częstościach wystąpień odpowiednie wartości średniej liczby wystąpień oraz średniego udziału procentowego wynoszą odpowiednio 529 razy i 0,1642% dla „PB" oraz 305 razy i 0,1657% dla „ZIN", co daje stosunek frekwencji na poziomie 1,7. Interesującym jest fakt, że średnia liczba wystąpień dla 100 najpopularniejszych leksemów „PB" jest również 1,7 razy większa niż dla odpowiednich leksemów „Zagadnień Informacji Naukowej". Natomiast różnica w średnim udziale procentowym dla 100 leksemów o najwyższych frekwencjach wynosi 0,0014%, czyli jest o rząd wielkości mniejsza niż dla 20 słów najpopularniejszych dla obu czasopism[200]. Z kolei stosunek frekwencji leksemów z obu zbiorów tekstów wynosi 1,42, czyli jest porównywalny z odpowiednimi wielkościami dla jednostek występujących najczęściej.

Interesujące wyniki dało porównanie najczęściej występujących leksemów dla obu czasopism. Wyraźnie zaznaczają się różnice w rozkładzie częstości (a w konsekwencji udziału procentowego w danym korpusie składowym) poszczególnych słów. Różnice te wynikają częściowo z odmiennych charakterystyk naukowych obu czasopism. Zjawisko rozprosze-

[200] Opracowanie własne na podstawie analizy artykułów wchodzących w skład korpusu badawczego.

nia frekwencji leksemów w obu omawianych częściach korpusu badawczego jest o tyle interesujące, że często w obu czasopismach publikowały te same osoby, oczywiście zamieszczane były różne artykuły.

Tabela 15. pozwala porównać rozkłady częstości wystąpień 20 leksemów o najwyższych frekwencjach w połączonych tekstach oraz w tekstach z poszczególnych czasopism. Analizując te dane można zauważyć, że udział procentowy wystąpień leksemu **biblioteka** jest najwyższy zarówno dla artykułów z poszczególnych periodyków, jak i w połączonym zbiorze tekstów. Drugim co do liczby wystąpień leksemem jest **informacja**, którego udział procentowy wzrósł o jedną pozycję w porównaniu z udziałami w poszczególnych zbiorach artykułów obu analizowanych tytułów.

Tabela 15. Porównanie częstości wystąpień dwudziestu najpopularniejszych leksemów w przetworzonym połączonym korpusie wraz z ich udziałami procentowymi w artykułach z „Przeglądu Bibliotecznego" i „Zagadnień Informacji Naukowej".

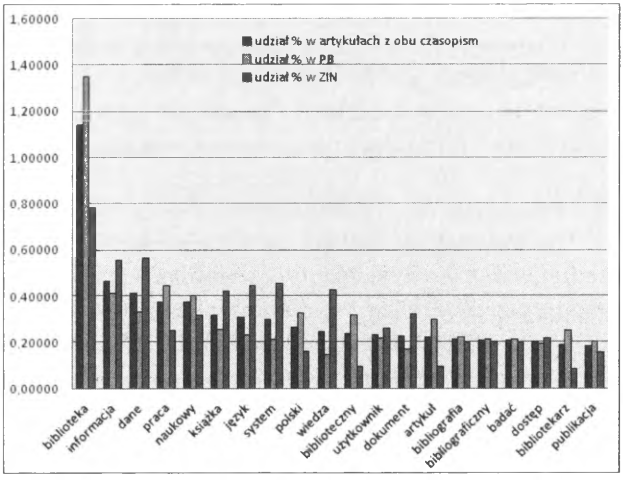
Lp.	Leksem	Udział procentowy w artykułach	Udział procentowy w „PB"	Udział procentowy w „ZIN"
1	biblioteka	1,144	1,350	0,785
2	informacja	0,462	0,412	0,550
3	dane	0,412	0,328	0,561
4	praca	0,373	0,443	0,249
5	naukowy	0,371	0,402	0,317
6	książka	0,315	0,255	0,421
7	język	0,307	0,229	0,444
8	system	0,298	0,210	0,452
9	polski	0,265	0,325	0,161
10	wiedza	0,247	0,147	0,423
11	biblioteczny	0,233	0,314	0,093
12	użytkownik	0,232	0,215	0,260
13	dokument	0,225	0,169	0,322
14	artykuł	0,223	0,296	0,095
15	bibliografia	0,212	0,219	0,198

Lp.	Leksem	Udział procentowy w artykułach	Udział procentowy w „PB”	Udział procentowy w „ZIN”
16	bibliograficzny	0,207	0,210	0,204
17	badać	0,206	0,211	0,198
18	dostęp	0,201	0,191	0,217
19	bibliotekarz	0,189	0,249	0,085
20	publikacja	0,184	0,201	0,155

Źródło: Opracowanie własne na podstawie analizy artykułów z „PB” i „ZIN” przez aplikację autorską.

Wykres 2. umożliwia porównanie udziałów procentowych poszczególnych leksemów w zbiorze połączonych tekstów z obu czasopism oraz udziału tych samych leksemów w treści artykułów w każdym z czasopism osobno. Uwzględniono dane o frekwencjach 20 słów najczęściej występujących w korpusie połączonym. Graficzna prezentacja tych danych pozwala wyraźnie zaobserwować zdecydowaną przewagę udziału procentowego leksemu **biblioteka** nad pozostałymi najczęściej pojawiającymi się w tekstach leksemami.

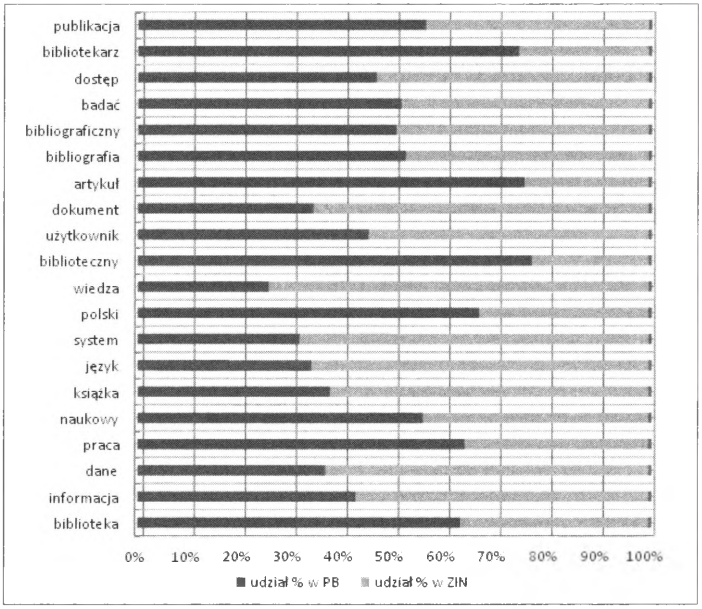
Wykres 2. Porównanie udziałów procentowych poszczególnych leksemów w zbiorze połączonych tekstów z „PB” i „ZIN” oraz w zbiorach artykułów z obu czasopism.



Źródło: Opracowanie własne na podstawie analizy artykułów z „PB” i „ZIN” przez aplikację autorską.

Z kolei wykres 3. odzwierciedla wzajemny stosunek częstości poszczególnych słów najczęściej występujących w zbiorze połączonych tekstów z obu czasopism w zbiorach tekstów z poszczególnych tytułów.

Wykres 3. Porównanie częstości wystąpień poszczególnych słów w artykułach z obu analizowanych czasopism.



Źródło: Opracowanie własne na podstawie analizy artykułów z „PB" i „ZIN" przez aplikację autorską.

Na podstawie analizy danych frekwencyjnych dla słownictwa z artykułów z obu czasopism fachowych można wysnuć wniosek, że słownictwo „Przeglądu Bibliotecznego" jest mniej różnorodne niż „Zagadnień Informacji Naukowej", natomiast duże częstości słów charakterystycznych pozwalają stwierdzić, że leksyka artykułów „PB" jest bardziej zwarta tematycznie.

4.1.2. Materiały konferencyjne

W kolejnej części rozdziału zostaną zaprezentowane wyniki analizy kwantytatywnej tekstów uzyskanych z publikowanych materiałów konferencyjnych z lat 2005-2008.

Również w przypadku artykułów konferencyjnych najczęściej występującym słowem jest **biblioteka**, która pojawia się w tym zbiorze 6062 razy, co daje udział procentowy na poziomie 1,77%. Tabela 16. umożliwia porównanie 20 słów występujących najczęściej w korpusie czasopism oraz w korpusie materiałów konferencyjnych. Także w tym zestawieniu, zamiast wartości bezwzględnych określających frekwencje poszczególnych słów, zaprezentowano udziały procentowe w odpowiednich częściach korpusu.

Tabela 16. Porównanie frekwencji najczęściej występujących słów w zbiorach tekstów z obu części składowych korpusu.

Artykuły z czasopism			Artykuły z materiałów konferencyjnych		
Lp.	**Leksem**	**Udział [%]**	**Lp.**	**Leksem**	**Udział [%]**
1	biblioteka	1,1443	1	biblioteka	1,7705
2	informacja	0,4620	2	użytkownik	0,4672
3	dane	0,4124	3	naukowy	0,4488
4	praca	0,3725	4	dane	0,4144
5	naukowy	0,3713	5	praca	0,3601
6	książka	0,3150	6	informacja	0,3355
7	język	0,3073	7	Internet	0,3329
8	system	0,2982	8	elektroniczny	0,3125
9	polski	0,2650	9	bibliotekarz	0,3081
10	wiedza	0,2471	10	formacja	0,2684
11	biblioteczny	0,2332	11	system	0,2581
12	użytkownik	0,2315	12	cyfrowy	0,2479
13	dokument	0,2247	13	baza	0,2453
14	artykuł	0,2232	14	biblioteczny	0,2348
15	bibliografia	0,2115	15	czasopismo	0,2330
16	bibliograficzny	0,2074	16	dostępny	0,2190
17	badać	0,2062	17	informacyjny	0,2096
18	dostęp	0,2006	18	zasób	0,2079

Artykuły z czasopism			Artykuły z materiałów konferencyjnych		
Lp.	Leksem	Udział [%]	Lp.	Leksem	Udział [%]
19	bibliotekarz	0,1890	19	wiedza	0,1956
20	publikacja	0,1844	20	katalog	0,1942

Źródło: Opracowanie własne na podstawie analizy artykułów materiałów konferencyjnych oraz z czasopism przez aplikację autorską.

Porównując dane frekwencyjne dotyczące dwudziestu najczęściej występujących słów w dwóch analizowanych zbiorach tekstów, można zauważyć, że najpopularniejsze słownictwo z obu zbiorów pokrywa się w 50%. Ponadto w obu zbiorach najczęściej występującym słowem jest **biblioteka**, przy czym udział tego leksemu jest wyższy w przypadku materiałów konferencyjnych i wynosi 1,77%. Różnie rozkładają się udziały procentowe poszczególnych leksemów w obu analizowanych zbiorach tekstów. W tekstach z materiałów konferencyjnych znacznie częściej niż w tekstach z czasopism występują leksemy **użytkownik** oraz **bibliotekarz** (oba leksemy zyskują aż 10 pozycji w rankingu częstości). Z kolei w zbiorze artykułów „ZIN" i „PB" zdecydowanie częściej pojawiają się słowa **wiedza** oraz **informacja**. Różnice w częstościach poszczególnych słów należących do strefy słownictwa charakterystycznego mogą wskazywać na zawężenie tematyczne tekstów z obu zbiorów.

Dla dwudziestu najczęstszych leksemów w zbiorze tekstów z czasopism oraz zbiorze tekstów z materiałów konferencyjnych średnia liczba wystąpień i średni procentowy udział wystąpień wynosi odpowiednio 1565 razy i 0,3090% dla periodyków oraz 1244 razy i 0,3632% dla drugiej części korpusu. Można zatem stwierdzić, że średni udział procentowy najpopularniejszych leksemów w tekstach w obu zbiorach jest porównywalny, natomiast średnia liczba wystąpień wskazuje na wyższą gęstość słownictwa strefy charakterystycznej w artykułach umieszczonych w czasopismach. Z kolei wśród 100 słów o najwyższych częstościach wystąpień odpowiednie wartości średniej liczby wystąpień oraz średniego udziału procentowego wynoszą odpowiednio 801 razy i 0,1583% dla periodyków oraz 545 razy i 0,1592% dla materiałów konferencyjnych, co daje stosunek frekwencji na poziomie 1,46. Natomiast stosunek frekwencji leksemów z obu zbiorów tekstów wynosi 1,42, czyli jest porównywalny z odpowiednimi wielkościami dla jednostek występujących naj-

częściej[201].

Ogólnie, w traktowanych niezależnie zbiorach tekstów z czasopism oraz z materiałów konferencyjnych 10.042 leksemy występujące w tekstach periodyków pojawiają się również w tekstach materiałów. Zatem słownictwo występujące w czasopismach w 46% pokrywa leksykę występującą w dokumentach publikowanych w materiałach konferencyjnych. Z drugiej zaś strony, 9902 leksemy z artykułów konferencyjnych pojawiają się w tekstach „PB" i „ZIN", co daje 36% pokrycia leksyki. Średnio w tekstach obu zbiorów pokrywają się ok. 9972 leksemy, czyli ok. 41% leksyki jest wspólne.

4.1.3. Analiza całego korpusu

Korpus tekstów fachowych utworzony na potrzeby niniejszej książkizawiera po wstępnym opracowaniu 848.650 słów, pochodzących ze 183 krótkich form piśmienniczych. Po sprowadzeniu wyrazów do postaci podstawowej korpus liczy 38.272 leksemy o średniej częstości ok. 22 wystąpień (22,17%) oraz średnim udziale procentowym leksemu w zbiorze tekstów z zakresu nauki o informacji kształtującym się na poziomie 0,0017%.

W całym korpusie 53 jednostki leksykalne cechują się częstościami wystąpień powyżej 1000, co stanowi 0,14% słownika, a jednocześnie suma wystąpień wszystkich form leksemów najczęstszych tworzy 16,7% zebranego tekstu.

Powyżej lub równo 100 razy (co I. Kamińska-Szmaj określa słownictwem bardzo częstym[202]) w tekście pojawia się 1049 wyrazów, co stanowi 2,74% słownika hasłowego badanego korpusu. Łącznie słowoformy bardzo częste tworzą 62% tekstu. W porównaniu do wyników opracowania danych statystycznych ze *Słownika frekwencyjnego*, podanych przez wspomnianą badaczkę, w korpusie tekstów z zakresu informacji naukowej wyrazy o wysokich frekwencjach (≥100) pojawiają się o 1,07 punktu procentowego częściej, niż w polszczyźnie pisanej, opracowanej w przytaczanym *Słowniku frekwencyjnym* (2,74% do 1,67%). Słownictwo bardzo częste stanowi w korpusie badawczym 62% tekstu (wobec 58% w przypadku polszczyzny współczesnej, co daje różnicę 4 punktów

[201] Opracowanie własne na podstawie analizy artykułów z korpusu badawczego.
[202] Za: I. Kamińska-Szmaj: dz. cyt., s. 20.

procentowych)[203]. Zatem w stosunku do porównywanej próby tekstów zbiorczych w dokumentach poświęconych nauce o informacji mamy do czynienia z wyższą koncentracją słownictwa charakterystycznego, przy zachowaniu średniego poziomu pokrycia tekstu przez wyrazy z tej strefy leksyki.

Z kolei w porównaniu do odpowiedniej części słownictwa tekstów popularnonaukowych opracowanych na potrzeby *Słownika frekwencyjnego*, jednostki bardzo częste w korpusie badawczym utworzonym na potrzeby niniejszej książki, stanowią aż o 27 punktów procentowych więcej tekstu (62% do 35%). Tak duży współczynnik pokrycia tekstów przez słownictwo bardzo częste jest charakterystyczny, w korpusie *Słownika frekwencyjnego...*, dla dramatu (odpowiednio wynosi 56,53%)[204]. Zatem w przypadku współczesnych (lata 2005-2008) tekstów naukowych z zakresu nauki o informacji mamy do czynienia ze znacznie wyższą koncentracją słownictwa bardzo częstego, niż miało to miejsce w przypadku stylu popularnonaukowego w polszczyźnie lat 60. XX wieku. Może to być spowodowane pojawieniem się wielu terminów specjalistycznych w terminologii fachowej informacji naukowej. Różnica ta może również wynikać z odmiennej metody doboru próby. Na potrzeby *Słownika frekwencyjnego* analizowane były losowe fragmenty tekstów z rożnych stylów funkcjonalnych piśmiennictwa, wielkość próbki wynosiła 50 wyrazów. Natomiast na potrzeby badawcze niniejszej książki analizowano pełne teksty z określonej dziedziny wiedzy.

Można by w tym miejscu postawić pytanie o zasadność porównywania wyników analiz współczesnego języka polskiego z odpowiednimi badaniami dla polszczyzny lat 60. XX wieku. Jak już wcześniej nadmieniono, język naturalny jest strukturą dynamiczną, odzwierciedlającą w słownictwie zmiany zachodzące w życiu ludzkim. Jednakże obecnie nie dysponujemy równie obszernymi opracowaniami frekwencyjnymi języka polskiego. Jak podaje PWN w witrynie *Korpusu języka polskiego*, osobiste przeliczanie częstości wybranych leksemów na potrzeby *Listy słów* potwierdziło, że rozkład lematów w tekstach starszych i współczesnych jest taki sam[205].

---

[203] Porównania wyników otrzymanych podczas prac badawczych przeprowadzonych na potrzeby niniejszej książki przeprowadzono w odniesieniu do wyników statystyk polszczyzny współczesnej oraz statystyk tekstów popularnonaukowych podanych w opracowaniu I. Kamińskiej-Szmaj: tamże, s. 20.

[204] Por. I. Kamińska-Szmaj: dz. cyt., s. 20.

[205] Za: *Lista słów*, dz. cyt.

Do strefy słownictwa częstego i średnio częstego ($10 \leq F < 100$) w badanym korpusie tekstów kwalifikuje się 5076 leksemów, co stanowi ok. 13% zbioru haseł. W wartościach bezwzględnych częstości wystąpień słownictwo tej strefy pokrywa 28% tekstu. W przypadku tak określonej granicy częstości, z danych ze *Słownika frekwencyjnego*, porównać można jedynie wskaźniki dla tekstów popularnonaukowych, ponieważ w dalszej części analizy i interpretacji w przytaczanej pracy zaliczone zostały leksemy o frekwencjach nie mniejszych niż 45%[206]. W tekstach popularnonaukowych polszczyzny lat 60. XX wieku wyrazy średnie i częste pokrywały 36% słownika. We współczesnych tekstach z zakresu nauki o informacji notujemy o 9 punktów procentowych mniej form wyrazowych o średnich i wysokich frekwencjach (27% do 36%).

Wyrazy rzadkie ($F < 10$) stanowią aż 84% słownika leksemów analizowanego zbioru, ale obejmują jedynie 10% tekstu. Natomiast poniżej średniej wystąpień (22 razy) pojawia się 90% słownika, co wyrażone frekwencjami bezwzględnymi daje pokrycie tekstu na poziomie ok. 19%.

Tabela 17. prezentuje rozkład leksemów korpusu badawczego w poszczególnych strefach słownictwa.

Tabela 17. Rozkład leksemów w strefach słownictwa korpusu głównego.

	Bardzo częste $F \geq 100$	**Częste i średnio częste $10 \leq F < 100$**	**Rzadkie $F < 10$**
Procent słownictwa	3%	13%	84%
Procentowe pokrycie tekstu	62%	28%	10%

Źródło: Opracowanie własne.

W tabeli 18. zaprezentowany jest rozkład frekwencyjny trzydziestu najpopularniejszych w całym korpusie leksemów.

[206] Za: I. Kamińska-Szmaj: dz. cyt., s. 23.

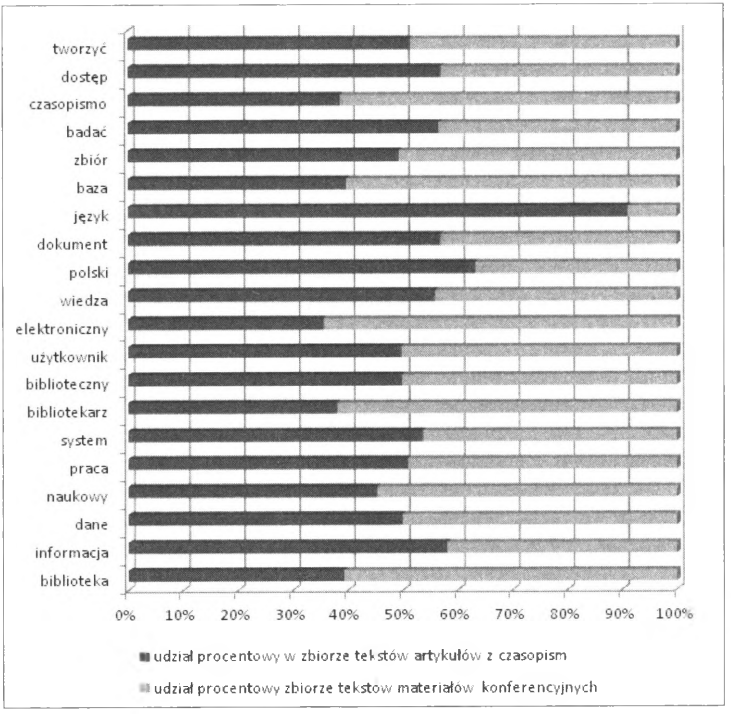
Tabela 18. Częstości wystąpień oraz udział procentowy trzydziestu najpopularniejszych leksemów w korpusie badawczym tekstów z zakresu informacji naukowej.

Lp.	Leksem	Liczba wystąpień	Udział procentowy w korpusie
1	biblioteka	11855	1,397
2	informacja	5597	0,660
3	dane	3507	0,413
4	naukowy	3417	0,403
5	praca	3119	0,368
6	system	2394	0,282
7	bibliotekarz	2012	0,237
8	biblioteczny	1985	0,234
9	użytkownik	1974	0,233
10	elektroniczny	1945	0,229
11	wiedza	1921	0,226
12	polski	1871	0,220
13	dokument	1722	0,203
14	język	1660	0,196
15	baza	1658	0,195
16	zbiór	1609	0,190
17	badać	1587	0,187
18	czasopismo	1540	0,181
19	dostęp	1535	0,181
20	tworzyć	1483	0,175
21	zasób	1408	0,166
22	dostępny	1402	0,165
23	informacyjny	1398	0,165
24	cyfrowy	1370	0,161
25	organizacja	1288	0,152
26	artykuł	1284	0,151

Lp.	Leksem	Liczba wystąpień	Udział procentowy w korpusie
27	bibliograficzny	1265	0,149
28	Internet	1256	0,148
29	publikacja	1236	0,146
30	nazywać	1229	0,145

Źródło: Opracowanie własne na podstawie analizy artykułów materiałów konferencyjnych oraz z czasopism przez aplikację autorską.

Wykres 4. Porównanie częstości wystąpień poszczególnych słów w artykułach z obu części korpusu badawczego.



Źródło: Opracowanie własne na podstawie analizy artykułów materiałów konferencyjnych oraz z czasopism przez aplikację autorską.

Podobnie jak miało to miejsce w obu zbiorach składowych korpusu tekstów, najwyższą frekwencją cechuje się słowo **biblioteka**, którego udział procentowy w korpusie wynosi ok. 1,4%. Kolejny najpopularniejszy leksem, **informacja**, pojawia się w tekstach zdecydowanie rzadziej, jego udział procentowy w zbiorze wynosi 0,46%. Pośród 30 najpopularniejszych w zbiorze słów aż 21 (czyli 70%) można uznać za słownictwo strefy charakterystycznej. Generalnie korpus badawczy stworzony na potrzeby niniejszej książki cechuje się dużym udziałem słownictwa charakterystycznego wśród najczęściej występujących leksemów, a warto przypomnieć, że właśnie ze strefy słownictwa charakterystycznego można najpewniej wskazać słowa kluczowe. Wykres 4. prezentuje rozkład udziałów procentowych 20 najpopularniejszych w całym korpusie badawczym tekstów w obu częściach składowych.

Wewnętrzne zróżnicowanie leksyki w korpusie badawczym pozwoli zobrazować tabela 19., prezentująca rozkład frekwencji 20 najpopularniejszych słów korpusu w obu jego częściach składowych.

Tabela 19. Porównanie częstości wystąpień dwudziestu najpopularniejszych leksemów w przetworzonym połączonym korpusie wraz z ich udziałami procentowymi w artykułach z „PB" i „ZIN" oraz w artykułach z materiałów konferencyjnych.

Lp.	Leksem	Liczba wystąpień w całym zbiorze	Udział procentowy w całym zbiorze	Udział procentowy w zbiorze tekstów artykułów z czasopism	Udział procentowy w zbiorze tekstów materiałów konferencyjnych
1	biblioteka	11855	1,397	1,144	1,770
2	informacja	5597	0,660	0,462	0,336
3	dane	3507	0,413	0,412	0,414
4	naukowy	3417	0,403	0,371	0,449
5	praca	3119	0,368	0,373	0,360
6	system	2394	0,282	0,298	0,258
7	bibliotekarz	2012	0,237	0,189	0,308
8	biblioteczny	1985	0,234	0,233	0,235
9	użytkownik	1974	0,233	0,232	0,234

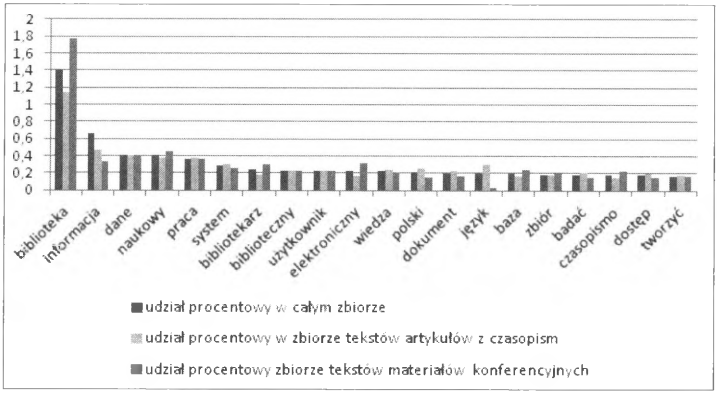
Lp.	Leksem	Liczba wystąpień w całym zbiorze	Udział procentowy w całym zbiorze	Udział procentowy w zbiorze tekstów artykułów z czasopism	Udział procentowy w zbiorze tekstów materiałów konferencyjnych
10	elektroniczny	1945	0,229	0,173	0,313
11	wiedza	1921	0,226	0,247	0,196
12	polski	1871	0,220	0,265	0,154
13	dokument	1722	0,203	0,225	0,171
14	język	1660	0,196	0,307	0,030
15	baza	1658	0,195	0,162	0,245
16	zbiór	1609	0,190	0,187	0,193
17	badać	1587	0,187	0,206	0,159
18	czasopismo	1540	0,181	0,147	0,233
19	dostęp	1535	0,181	0,201	0,152
20	tworzyć	1483	0,175	0,179	0,169

Źródło: Opracowanie własne na podstawie analizy artykułów z materiałów konferencyjnych oraz z czasopism przez aplikację autorską.

Zestawienie danych z tabeli łatwiej przeanalizować, gdy dane zaprezentowane są w postaci graficznej, jak na wykresie 5.

Wykres obrazuje wyraźną przewagę udziału procentowego leksemu **biblioteka** nad pozostałymi leksemami. Wartości określające liczbę wystąpień tego leksemu w zbiorze analizowanych tekstów mogą sugerować przynależność wyrazu biblioteka do strefy słownictwa częstego. Jednakże zgodnie z przyjętą w niniejszej książce definicją, ze względu na swoją funkcję gramatyczną wyraz ten zaliczamy do strefy słownictwa charakterystycznego. W wielokrotnie cytowanym *Korpusie Języka Polskiego PWN* na liście frekwencyjnej wśród 200 najpopularniejszych leksemów języka polskiego nie występuje słowo **biblioteka**, co może stanowić dodatkową przesłankę, że przypisanie tego wyrazu do strefy słownictwa charakterystycznego jest decyzją słuszną, szczególnie po uwzględnieniu zakresu tematycznego analizowanych dokumentów.

Wykres 5. Porównanie udziałów procentowych poszczególnych leksemów w całym korpusie oraz w jego częściach: czasopiśmienniczej i konferencyjnej.



Źródło: Opracowanie własne na podstawie analizy artykułów z materiałów konferencyjnych oraz z czasopism przez aplikację autorską.

4.2. Analiza korpusu pomocniczego

Korpus kontrolny tekstów z zakresu ekonomii i zarządzania, utworzony na potrzeby niniejszej książki, zawiera po wstępnym opracowaniu 195.300 tokenów. Po sprowadzeniu wyrazów do postaci podstawowej słownik korpusu liczy 10.228 leksemów o średniej częstości ok. 19 wystąpień (dokładnie 19,09).

Do strefy słownictwa bardzo częstego ($F \geq 100$) należy 195 leksemów, co stanowi 1,91% słownika. Łącznie słowoformy bardzo częste tworzą 33% tekstu. W przypadku tekstów ekonomicznych wyrazy bardzo częste stanowią porównywalną część słownika, jak ma to miejsce w polszczyźnie pisanej (odpowiednio 1,91% i 1,67%), różnica ta jest mniejsza niż w przypadku porównania tekstów z zakresu informacji naukowej z polszczyzną pisaną[207]. Pokrycie tekstu przez leksemy bardzo częste w badanym zbiorze jest z kolei znacznie niższe niż dla tekstów ogólnych języka polskiego (33% do 58%) i jest zbliżone do odpowiedniej wartości dla tekstów popularnonaukowych (35,4%)[208].

[207] Niższy jest też stopień pokrycia tekstu przez słownictwo bardzo częste (33% w przypadku nauk ekonomicznych do 62% w przypadku informacji naukowej). Może to świadczyć o szybszym przyroście terminologii fachowej w leksyce nauki o informacji niż w leksyce nauk ekonomicznych.

[208] Por. I Kamińska-Szmaj: dz. cyt., s. 20.

Do strefy słownictwa częstego i średnio częstego ($10 \leq F < 100$) w omawianym kontrolnym korpusie tekstów kwalifikuje się 1908 leksemów, co stanowi ok. 18% zbioru haseł. W wartościach bezwzględnych częstości wystąpień słownictwo tej strefy pokrywa 49% tekstu. W porównaniu do statystyk dotyczących tekstów popularnonaukowych ze *Słownika frekwencyjnego* jest to znaczna różnica – 49% do 36%, co daje różnicę 13 punktów procentowych.

Wyrazy rzadkie ($F < 10$) stanowią aż 79% słownika leksemów analizowanego zbioru, a jedynie 18% tekstu. Natomiast poniżej średniej wystąpień (19 razy) pojawia się ok. 88% haseł słownika, co jest wielkością porównywalną do odpowiednich wyliczeń dla tekstów z zakresu informacji naukowej, natomiast w frekwencji bezwzględnej daje pokrycie tekstu na poziomie ok. 30%, czyli ok. 1,5 razy więcej niż w przypadku słownictwa z głównego korpusu badawczego (11 punktów procentowych różnicy).

Tabela 20. prezentuje rozkład leksemów korpusu badawczego w poszczególnych strefach słownictwa.

Tabela 20. Rozkład leksemów w strefach słownictwa korpusu głównego, kontrolnego oraz tekstów popularnonaukowych.

Słownictwo	**Bardzo częste F ≥100**			**Częste i średnio częste 10 ≤ F < 100**			**Rzadkie F < 10**		
Korpus	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Procent słownictwa	3%	2%	2%	13%	18%	b.d.	84%	80%	b.d.
Procent pokrycia tekstu	62%	33%	35%	28%	49%	36%	10%	18%	29%

Źródło: Opracowanie własne.

1 – Informacja naukowa,
2 – Ekonomia i zarządzanie,
3 – Teksty popularnonaukowe[209].

Natomiast w tabeli 21. zaprezentowany jest rozkład frekwencyjny trzydziestu leksemów najpopularniejszych w korpusie kontrolnym.

[209] Dla tekstów popularnonaukowych dane cyt. za: tamże, s. 20.

Tabela 21. Częstości wystąpień oraz udział procentowy trzydziestu najpopularniejszych leksemów w korpusie badawczym tekstów z zakresu ekonomii i zarządzania.

Lp.	Leksem	Liczba wystąpień	Udział procentowy w korpusie
1	model	1320	1,175
2	badać	742	0,661
3	wartość	667	0,594
4	przedsiębiorstwo	649	0,578
5	cena	573	0,510
6	praca	554	0,493
7	proces	479	0,426
8	wynik	457	0,407
9	stosować	449	0,400
10	nowy	433	0,385
11	metoda	414	0,369
12	analiza	411	0,366
13	rynek	390	0,347
14	poziom	383	0,341
15	tabela	377	0,336
16	źródło	368	0,328
17	stać	351	0,312
18	przypadek	351	0,312
19	zmienna	346	0,308
20	liczba	324	0,288
21	podstawa	320	0,285
22	własny	319	0,284
23	prowadzić	315	0,280
24	działalność	311	0,277
25	szereg	307	0,273
26	okres	305	0,272

Lp.	Leksem	Liczba wystąpień	Udział procentowy w korpusie
27	ekonomiczny	295	0,263
28	prognoza	285	0,254
29	zmiana	283	0,252
30	wzrost	281	0,250

Źródło: Opracowanie własne.

W zbiorze tekstów ekonomicznych najwyższą frekwencją cechuje się słowo **model**, którego udział procentowy w korpusie wynosi ok. 1,2%. Kolejny najpopularniejszy leksem, **badać**, pojawia się w tekstach zdecydowanie rzadziej, jego udział procentowy w zbiorze wynosi 0,66%. Pośród 30 najpopularniejszych w zbiorze słów 17 (czyli 57%) można uznać za słownictwo strefy charakterystycznej.

4.3. Słowa kluczowe uzyskane w wyniku indeksowania tradycyjnego i automatycznego

Opisane operacje przetwarzania tekstów języka naturalnego zmierzały do ustalenia frekwencji poszczególnych leksemów, a następnie wyznaczenia ich wagi dla danego tekstu. Jak już wspomniano wcześniej, wykluczenie z analizowanych tekstów słów o najwyższych frekwencjach w języku polskim spowodowało, że najwyższe pozycje na listach rankingowych zajmują przeważnie wyrazy należące do zbioru słownictwa charakterystycznego dla danej dziedziny. Właśnie one stanowią zestaw leksyki, z którego pochodzą słowa kluczowe. W dalszej części rozdziału zostaną opisane wyniki zastosowania wybranych sposobów ustalania wagi leksemów oraz wskazywania słów kluczowych charakteryzujących treść artykułów.

Indeksatorzy ustalający słowa kluczowe nie byli ograniczeni liczbą wyrażeń lub słów, które mogą wskazać, w związku z czym w udzielonych przez nich odpowiedziach pojawiło się więcej słów kluczowych, niż podawali autorzy poszczególnych artykułów. Ponadto, ponieważ dla niektórych analizowanych artykułów uzyskano kilka niezależnych zestawów słów kluczowych, ostateczny ich zbiór stanowi sumę logiczną odpowiedzi podanych przez indeksatorów. W zbiorze będącym sumą kilku zesta-

wów przygotowanych tradycyjnie, każde wyrażenie występuje tylko raz, niemniej w ten sposób wzrosła redundancja w stosunku do zbiorów słów kluczowych ustalonych przez autorów analizowanych artykułów. W przypadku korpusu tekstów z zakresu informacji naukowej nadmiarowość wyrażeń wskazanych jako kluczowe przez indeksatorów w stosunku do liczby wyrażeń kluczowych wybranych przez autorów wynosi ok. 167% (średnio 7 wyrażeń więcej). Jako przykład nadmiarowości słów kluczowych wskazanych przez studentów uczestniczących w badaniach w stosunku do zestawów autorskich, może służyć następujący tekst, z którego zacytowano abstrakt, a następnie w tabeli 21. zaprezentowano dwa zestawy słów kluczowych, przygotowanych przez indeksatorów.

Abstrakt:

„W okresie rozwoju społeczeństwa informacyjnego biblioteka naukowa zyskuje w środowisku akademickim rolę istotnego ośrodka komunikacji pomiędzy użytkownikiem a jego potrzebami informacyjnymi. Jedną z form realizacji tych potrzeb jest usprawnienie procesu dostarczania poszukiwanych przez użytkownika informacji, m.in. przez budowanie bibliotek cyfrowych. Problem tworzenia i zarządzania zasobami elektronicznymi, możliwości ich wspólnego przeszukiwania oraz współpracy między bibliotekami w kontekście bieżącej informacji w naukach technicznych, został przedstawiony na przykładzie analizy zawartości polskich bibliotek cyfrowych, opartych na oprogramowaniu dLibra. W artykule przedstawiono wyniki badań zasobu cyfrowego, liczącego ponad 158.000 obiektów, wspólnie przeszukiwanego w ramach Federacji Bibliotek Cyfrowych. Na podstawie analizy zebranego materiału statystycznego, spróbowano ocenić zarówno wartość publikacji cyfrowych pod względem chronologicznym, jak i warsztat wyszukiwawczy. Najistotniejszym wnioskiem wynikającym z przeprowadzonych badań jest pilna konieczność ujednolicenia metadanych na poziomie każdej biblioteki i wszystkich bibliotek FBC, w celu wspólnego efektywnego wyszukiwania”[210].

[210] A. Kazan, E. Skubała: *Polskie biblioteki cyfrowe na platformie dLibra – zasób w kontekście tworzenia nowoczesnych kolekcji źródeł informacji dla nauk technicznych*. W: *Informacja dla nauki a świat zasobów cyfrowych*, pod red. H. Ganińskiej. Poznań: Biblioteka Główna Politechniki Poznańskiej 2008, s. 21.

Tabela 22. Zestawy słów kluczowych przygotowane przez indeksatorów dla przykładowego tekstu.

Słowa kluczowe autorskie	Słowa kluczowe wskazane przez indeksatorów
biblioteki cyfrowe; program dLibra; warsztat wyszukiwawczy	biblioteki cyfrowe; digitalizacja; dokumenty naukowe; federacja bibliotek cyfrowych; informacja naukowa; informacja; metadane; nauki techniczne; nowoczesne źródła informacji; platforma dLibra; polskie biblioteki cyfrowe; proces dostarczania informacji; społeczeństwo informacyjne; środowisko akademickie; systemy biblioteczne; technologie informacyjne; udostępnianie; warsztat wyszukiwawczy; współpraca bibliotek cyfrowych; zarządzanie zasobami elektronicznymi; źródła elektroniczne; źródła informacji

Źródło: Opracowanie własne na podstawie analizy tekstów.

Dla porównania typowo leksykalnego oba analizowane zestawy słów kluczowych, wskazanych w wyniku indeksowania tradycyjnego, zostały poddane automatycznym operacjom przetwarzania, takim jak pełne teksty. Dla przypomnienia, były to operacje usunięcia słów mało znaczących oraz ujednolicenia formy graficznej poszczególnych wyrazów (lematyzacja). Leksemy pochodzące z analizowanych zestawów słów kluczowych ustalanych w procesie tradycyjnym były następnie porównywane z leksemami kluczowymi generowanymi automatycznie. W przypadku zbiorów słów wskazanych przez człowieka zaprezentowanych w postaci list leksemów, listy lematów uzyskane od indeksatorów prezentowały liczniejsze słownictwo niż lematy pochodzące z zestawów od autorów artykułów. W przypadku korpusu tekstów z zakresu nauki o informacji, dla leksemów pochodzących z zestawów autorskich średnia liczba lematów w zbiorze wynosi 7,85, podczas gdy dla leksemów uzyskanych od studentów wartość ta równa się 17,35 leksemu. Indeksatorzy zaproponowali średnio 9,5 leksemu charakteryzującego treść artykułu, więcej niż autorzy artykułów. Odpowiednie wartości dla korpusu tekstów ekonomicznych kształ-

towały się na podobnym poziomie. Średnia liczba leksemów w wyrażeniach kluczowych użytych przez autorów wynosi 7,7, zaś indeksatorzy podali średnio 16 lematów, co daje statystycznie 8 haseł kluczowych więcej niż podanych przez autorów.

Analiza zestawów słów kluczowych wskazanych w wyniku indeksowania tradycyjnego wykazuje spore różnice pomiędzy wyrażeniami charakteryzującymi treść artykułów ustalanymi przez ich autorów a wskazywanymi przez indeksatorów. Pierwsza zauważalna różnica plegająca na liczbie wyrażeń, a w konsekwencji i leksemów, i słów kluczowych wskazuje, że osoby indeksujące charakteryzowały treść analizowanych tekstów bardziej szczegółowo niż autorzy owych tekstów. Różnica ta może wynikać z faktu, że autorzy poszczególnych artykułów mieli jeszcze do dyspozycji abstrakt jako dodatkową możliwość scharakteryzowania treści danego tekstu, tak więc nie potrzebowali odzwierciedlać jej przy użyciu słów kluczowych w sposób bardzo szczegółowy. Studenci zaś mieli do dyspozycji tylko słowa kluczowe jako sposób scharakteryzowania treści, stąd też zastosowali większą liczbę opisów. Prawdopodobnie autorzy dobierając słowa kluczowe dopasowywali je również od docelowej grupy odbiorców lub do charakteru publikacji, w której miały pojawić się artykuły. Ponadto liczba zaproponowanych wyrażeń kluczowych może być narzucona autorom przez zespoły redakcyjne zbiorów tekstów jako formalny ogranicznik liczby wyrażeń charakteryzujących treść danego tekstu.

Warto zauważyć, że liczba 7,85 (i odpowiednio 7,7) mieści się w zakresie pojemności ludzkiej pamięci krótkotrwałej, co ułatwia odbiorcy zapamiętanie każdego z nich oraz szybką ocenę treści danego tekstu. **Pamięć krótkotrwała** (ang. *short term memory*, STM) ma pojemność ograniczoną do 7 (± 2) jednostek[211]. Ponadto pojemność pamięci krótkotrwałej nie jest wartością stałą, ponieważ może dotyczyć różnych kategorii jednostek od liter, poprzez wyrazy, aż do wyrażeń złożonych[212]. Należy też pamiętać, że jednostka pamięci jest jednostką znaczeniową a nie formalną. Jak wskazują badania zakresu pamięci bezpośredniej, ich uczestnicy zapamię-

[211] Za: B. Sosińska-Kalata: *Modele organizacji wiedzy*..., s. 42.

[212] Za: I. Kurcz, A. Polkowska: *Interakcyjne i autonomiczne przetwarzanie informacji językowych. Na przykładzie procesu rozumienia tekstu czytanego na głos.* Wrocław: Zakład Narodowy im. Ossolińskich 1990, s. 19. W pracy tej można znaleźć również dyskusję rozróżnienia pomiędzy funkcjonowaniem pamięci krótko- a długotrwałej (tamże, s. 20-23), jednakże na potrzeby niniejszej książki wystarczy przyjąć, że w pamięci krótkoterminowej przechowujemy 7±2 jednostki oraz że mogą to być słowa lub wyrażenia kluczowe.

tywali jednakową liczbę elementów niezależnie od ich rodzaju, tzn. tyle samo pojedynczych słów, jak i wyrażeń złożonych z kilku słów[213].

Wydawać by się mogło, że znaczna nadmiarowość leksemów w wyrażeniach kluczowych wskazanych przez indeksatorów (221% więcej leksemów niż w zestawach autorskich dla informacji naukowej i 204% dla ekonomii i zarządzania) zapewni całkowitą zgodność pomiędzy dwoma zestawami słów kluczowych ustalanych tradycyjnie. Rzeczywiście dla niektórych artykułów z zakresu nauki o informacji zgodność lematów kluczowych wybranych przez indeksatorów w stosunku do lematów wybranych przez autorów wynosiła właśnie 100%, jednakże dla niektórych jedynie 20%, co jest najniższą odnotowaną w trakcie badań wartością podobieństwa. Pełną zgodność odnotowano jedynie w przypadku 45% analizowanych zestawów słów kluczowych poddanych optymalizacji lingwistycznej. Średnia zgodność zestawów słów kluczowych wskazanych przez człowieka i wyrażonych w postaci list leksemów wynosi 77%. W przypadku zaś tekstów z zakresu nauk ekonomicznych i zarządzania można stwierdzić, że najniższa wartość zgodności pomiędzy dwoma zestawami słów kluczowych wskazanych tradycyjnie wynosiła 33%. W przypadku tego zbioru tekstów średnia zgodność zestawów słów kluczowych ustalanych przez człowieka wynosi 70%, czyli jest porównywalna z odpowiednim współczynnikiem dla korpusu tekstów z zakresu informacji naukowej.

W tabeli 23. zaprezentowano porównanie przykładowych zestawów słów kluczowych wskazanych przez autorów artykułów oraz przez indeksatorów dla tekstów z zakresu nauki o informacji. W celu ułatwienia porównań, wyrażenia kluczowe zostały posortowane alfabetycznie, nie zachowano porządku zaproponowanego przez ich autorów.

Analiza podobieństw pomiędzy dwoma zestawami słów kluczowych wskazanymi dla jednego tekstu (zestaw autorski i zestaw studencki) w przypadku obu korpusów tekstów przygotowanych na potrzeby niniejszej książki pozwala stwierdzić, że indeksatorzy wskazywali wyrażenia kluczowe z dużym podobieństwem do wyrażeń zaproponowanych przez autorów analizowanych artykułów, jednakże stopień zgodności był niższy niż można by przypuszczać.

---

[213] I. Kurcz: *Pamięć*. W: *Pamięć. Uczenie się. Język*, pod red. T. Tomaszewskiego. Warszawa: Wydawnictwo Naukowe PWN 1995, s. 9.

Tabela 23. Porównanie zestawów słów kluczowych przypisanych do przykładowych artykułów przez ich autorów ze słowami kluczowymi wskazanymi przez indeksatorów.

Tytuł	Autorskie słowa kluczowe	Słowa kluczowe wskazane przez indeksatorów
BazTOL jako przykład serwisu typu subject gateway o kontrolowanej jakości	baza danych; nauki techniczne; portal; serwis tematyczny o kontrolowanej jakości; wyszukiwanie informacji w Internecie	baza danych; nauki techniczne; katalogowanie zasobów sieciowych; portal; serwis tematyczny o kontrolowanej jakości; struktura serwisu; wyszukiwanie informacji w Internecie
Biblioteka akademicka jako element globalnej cyfrowej infrastruktury informacyjnej	biblioteki akademickie; globalna infrastruktura informacyjna	biblioteki akademickie; biblioteki cyfrowe; cyfrowa infrastruktura informacyjna; digitalizacja; formy kształcenia; globalizacja; globalna infrastruktura informacyjna; Internet; katalogowanie; kształcenie bibliotekarzy; nauczanie zdalne; organizacja przestrzeni bibliotecznej; rola bibliotek w nauczaniu ustawicznym
Biblioteka Jagiellońska 21. wieku – tradycja i nowoczesność	Biblioteka Jagiellońska; budownictwo biblioteczne; komputeryzacja bibliotek naukowych; NUKAT; współpraca bibliotek; zbiory biblioteczne - ochrona	Biblioteka Jagiellońska; digitalizacja zbiorów; informacja naukowa; komputeryzacja bibliotek; ochrona zbiorów
Bibliotekarze dyplomowani – wczoraj, dziś i jutro	bibliotekarstwo - Polska; bibliotekarze - status prawny	bibliotekarze w Polsce; bibliotekarze dyplomowani

Tytuł	Autorskie słowa kluczowe	Słowa kluczowe wskazane przez indeksatorów
Biblioteki akademickie – trendy dotyczące zasobów elektronicznych	archiwizacja Web; archiwizacja; czasopisma wolnodostępne; digitalizacja; kolekcje elektroniczne; konserwacja; repozytorium instytucjonalne; usługi informacyjne	archiwizacja informacji; biblioteki akademickie; biblioteki cyfrowe; digitalizacja; e-biblioteki; e-czasopisma; e-kolekcje; elektroniczne repozytoria; e-zbiory; konserwacja zbiorów; naukowe źródła informacji; technologie cyfrowe; technologie informacyjne; usługi informacyjne; zasoby elektroniczne; zasoby informacyjne; źródła informacji
Biblioteki naukowe wobec kulturowych i cywilizacyjnych potrzeb społeczeństwa	biblioteka hybrydowa; biblioteka naukowa; centrum edukacji, informacji i kultury; Internet	biblioteka hybrydowa; biblioteka naukowa; dziedzictwo kulturowe; edukacja; Internet; kultura; nauka; społeczeństwo informacyjne; technologie informacyjno-komunikacyjne
Elektroniczna książka na elektronicznym papierze. Czy to już zmierzch ery druku?	badanie użytkowników; e-czytnik; e-ink; elektroniczny papier; e-papier	badania ankietowe użytkowników; e-czytnik; e-papier; książka elektroniczna; papier elektroniczny
Od witryny internetowej do portalu bibliotecznego	metadane; portal biblioteczny; technologia informacji; zasoby internetowe	biblioteczne portale internetowe; Biblioteka Główna Politechniki Poznańskiej; dostęp do zasobów informacji; informacja naukowa; Internet; metadane; portal biblioteczny; rywalizacja dostawców informacji; technologia informacyjna; witryna internetowa; zadania portalu bibliotecznego; zasoby internetowe; zasób elektroniczny

Tytuł	Autorskie słowa kluczowe	Słowa kluczowe wskazane przez indeksatorów
Otwarte zasoby wiedzy na stronach internetowych wybranych bibliotek uczelni technicznych w Polsce i na świecie – przegląd z perspektywy doświadczeń Wypożyczalni Międzybibliotecznej Biblioteki Głównej Politechniki Warszawskiej	Biblioteka Główna Politechniki Warszawskiej; Internet; otwarte zasoby wiedzy; wypożyczanie międzybiblioteczne	Biblioteka Główna Politechniki Warszawskiej; Internet; metody wyszukiwania dokumentów; otwarte zasoby cyfrowe; otwarte zasoby wiedzy; procedury wyszukiwawcze; strony internetowe bibliotek; wypożyczanie międzybiblioteczne; zasoby naukowe
Podlaska Biblioteka Cyfrowa	biblioteka cyfrowa; digitalizacja	biblioteki cyfrowe; digitalizacja; Podlaska Biblioteka Cyfrowa; udostępnianie zbiorów; wdrażanie technologii informacyjnych; zadania biblioteki

Źródło: Opracowanie własne na podstawie analizy przykładowych artykułów.

4.3.1. Waga słów wyróżnionych w tekście

Jedną z hipotez badawczych niniejszej książki było sprawdzenie, czy wyróżnienie słowa przez autorów w treści tekstu dokumentu pozwala na zwiększenie wagi odpowiedniego leksemu na liście frekwencyjnej leksemów. Dla przypomnienia – jako wyróżnione traktowane były wszystkie słowa z odpowiednim formatowaniem tekstu w artykułach oryginalnych. Przy wskazywaniu takich wyrażeń uwzględniano następujące elementy formatowania: pogrubienie czcionki, zastosowanie stylu nagłówkowego. Wyjątkiem były wyrazy **bibliografia**, **literatura**, **podsumowanie**, **wstęp** oraz **zakończenie**, potraktowane w tym przypadku jako słowa nieznaczące, ponieważ jako elementy obowiązkowe aparatu naukowego, pojawiały się w każdym artykule. Wszystkie wyrazy wyróżnione odpowiednim formatowaniem, wyodrębnione z analizowanych tekstów, poddane zostały takim samym operacjom przetwarzania lingwistycznego, jakim podlegały pełne teksty. Uzyskane w wyniku przygotowania leksemy zostały następnie porównane z lekse-

mami reprezentującymi słowa kluczowe wskazane przez autorów dla poszczególnych artykułów. W przypadku tego porównania uzyskane wartości wahały się w zakresie od 0 do 100%. Całkowity brak dopasowania słów wyróżnionych do słów kluczowych w danym artykule wystąpił w przypadku 10% tekstów, zaś pełna zgodność jedynie w przypadku 5% tekstów (wartości te były zbliżone do siebie w przypadku obu korpusów tekstów). Średni stopień zgodności słów wyróżnionych formatowaniem ze słowami kluczowymi dla analizowanych dokumentów wynosi 33% zarówno w przypadku tekstów z zakresu informacji naukowej, jak i z zakresu ekonomii i zarządzania. Z kolei średni stopień podobieństwa słownictwa wyróżnionego z leksemami występującymi najczęściej w przetworzonych lingwistycznie tekstach wynosi 14%, co jest wartością zbyt małą, żeby można było na jej podstawie wiarygodnie wskazać słowa silnie związane z treścią tekstu[214].

Uwzględniając powyższe rezultaty należy stwierdzić, że słowa wyróżnione formatowaniem w dokumencie nie są w wystarczającym stopniu znaczące ani związane z treścią, żeby w istotny sposób zwiększały wagę leksemów branych pod uwagę jako słowa kluczowe. Można stąd wyciągnąć praktyczny wniosek, że podczas przetwarzania tekstów języka naturalnego nie ma potrzeby ponoszenia dodatkowych kosztów (czasowych i operacyjnych) na szczególne traktowanie słów wyróżnionych w tekście.

4.3.2. Słowa kluczowe wskazywane automatycznie

Kolejny etap procedury badawczej dotyczył możliwości automatycznego generowania słów kluczowych. Jak już wspomniano wcześniej, wykluczenie na wstępnym etapie wyrazów o małej wartości informacyjnej spowodowało, że najwyższe pozycje rankingowe, ze względu na częstość występowania, zajmowały leksemy należące do grupy słownictwa charakterystycznego. W związku z tym, że wynikiem analizy automatycznej były pojedyncze leksemy, w dalszych operacjach porównawczych również zestawy autorskich słów kluczowych ustalonych dla poszczególnych artykułów prezentowane były w postaci posortowanych alfabetycznie list leksemów. W zależności od przyjętej metody ważenia leksemów różnie kształtowała się zgodność słów kluczowych generowanych automatycznie z podanymi przez autorów.

[214] Przy założeniu, że uwzględnia się 20 najczęstszych leksemów z listy rankingowej, gdy zakres się zwiększa zgodność wzrasta – aż do 100% dla pełnych list.

Wybór leksemów jako formy słów kluczowych umożliwia swobodne dopasowanie charakterystyk wyszukiwawczych do kwerend użytkowników. Podczas wyszukiwania dokumentów na podstawie zapytania, wyrażenia przekazane przez użytkownika zostają sprowadzone do postaci kanonicznej i dopiero wtedy porównane z opisami treści dokumentów. Metoda ta pozwala zwiększyć relewancję odpowiedzi udzielonej przez system. Użytkownicy w swoich kwerendach podają zazwyczaj słowa kluczowe bez kontroli formalnej, najczęściej są to formy podstawowe, ale nierzadko różne inne słowoformy danego wyrazu. Ze względu na ograniczenia czasowo-operacyjne podczas badań przeprowadzonych na potrzeby niniejszej książki nie uwzględniano wyrażeń kluczowych złożonych (kilkuwyrazowych). Natomiast wydaje się, że jest to oczywisty kierunek dalszych badań – jak niejednokrotnie zostało to podkreślone, mocą indeksowania poprzez słowa kluczowe jest możliwość swobodnego łączenia ich w celu zawężenia zbioru wyników. Łączenie takie można przeprowadzić za pomocą wyrażeń algebry Boole'a również dla wyrażeń jednowyrazowych, ale możliwość operowania hasłami kluczowymi wielowyrazowymi usprawnia proces wyszukiwania informacji.

W przypadku słów wskazanych jako kluczowe automatycznie, pojawia się problem ograniczenia zbioru uzyskanych wyników. Ze względu na brak jakichkolwiek naturalnych wskaźników limitujących rozmiar zbioru potencjalnych słów kluczowych wygenerowanych automatycznie, postanowiono ustalić granicę listy: 17 pozycji dla tekstów z zakresu nauki o informacji oraz 15 pozycji dla tekstów ekonomicznych. Obie przyjęte wartości graniczne odpowiadają średniej liczbie leksemów wskazanych przez osoby indeksujące.

Dla zwykłego sortowania ze względu na liczbę wystąpień w danym tekście uzyskano średnio 54% pokrycia leksemów ze zbiorów autorskich słów kluczowych przez słowa wskazane automatycznie dla tekstów z zakresu informacji naukowej oraz 54,2 % dla tekstów ekonomicznych. Przykładowy rozkład lematów z obu porównywanych zbiorów dla głównego korpusu tekstów prezentuje tabela 24. Leksemy uzyskane w wyniku operacji automatycznych są wyświetlone w kolejności malejącej według liczby wystąpień.

Tabela 24. Porównanie słów kluczowych autorskich i wygenerowanych automatycznie dla przykładowego tekstu.

Słowa kluczowe autorskie	Słowa wygenerowane automatycznie
biblioteczny; informacja; internetowy; metadane; portal; technologia; zasób	portal; biblioteczny; biblioteka; zasób; usługa; użytkownik; informacja; dostęp; online; katalog; baza; wyszukiwać; system; *management*; internetowy; dokument; zarządzać

Źródło: Opracowanie własne na podstawie analizy przykładowych artykułów.

W przypadku kryterium liczby wystąpień w tekście, leksemy kluczowe wskazane automatycznie są o 23 punkty procentowe mniej zgodne z leksemami autorskimi niż leksemy wskazane przez studentów (77% do 54%). Natomiast średnia zgodność leksemów podanych przez indeksatorów z leksemami wskazanymi automatycznie wynosi w tym przypadku 52%.

Z kolei, gdy za kryterium przyjęto liczbę dokumentów, w których dany leksem się pojawił, zgodność w określonym zbiorze 17-elementowym spadła do 43%. Następnie uzyskany zbiór wyników dla każdego analizowanego tekstu został posortowany według malejących wartości IDF, obliczonych zgodnie ze wzorem:

$$idf_t = log_2 \frac{N}{df_t}$$

W tym przypadku zgodność pomiędzy dwoma zbiorami słów kluczowych – autorskim oraz automatycznie generowanym – spadła do 35%, natomiast dla zbiorów studenckiego i automatycznego poziom zgodności wynosi 44%.

Ostatnim badanym sposobem ważenia leksemów była miara tf-idf, dla której uzyskano średnią zgodność na poziomie 48%, zaś dla leksemów wskazanych przez studentów oraz automatycznie – 49%.

4.4. Ocena zastosowanych metod ustalania wagi słowa

Okazało się, że najskuteczniejszą metodą, dającą najwyższy współczynnik pokrycia leksemów wybranych przez autorów poszczególnych dokumentów przez leksemy wskazane w wyniku automatycznego przetwarzania tekstów, jest usunięcie słów o małej wartości informacyjnej w początkowej fazie procesu przetwarzania oraz posortowanie pozostałych leksemów w kolejności malejącej ze względu na częstość występowania w tekstach. Porównanie list słów kluczowych generowanych automatycznie umożliwia tabela 25. W tabeli przy poszczególnych metodach ważenia leksemów podano stopień zgodności z listą autorską, wyróżniono wyrazy zgodne.

Tabela 25. Zestawienie list słów kluczowych – autorskich i automatycznych, generowanych według różnych kryteriów oceny wagi leksemu.

Słowa kluczowe autorskie	**Słowa kluczowe generowane automatycznie, sortowane według:**			
	frekwencji (średnio 54%)	**liczby dokumentów** (średnio 43%)	**wartości IDF** (średnio 35%)	**wartości tf-idf** (średnio 48%)
baza	zasób	opis	Baztol	Baztol
informacja	redaktor	**informacja**	edycja	zasób
Internet	źródło	redaktor	**portal**	katalogować
jakość	Baztol	dostęp	kryterium	rekord
kontrolować	katalogować	tworzyć	rekord	**portal**
nauka	**Internet**	**baza**	katalogować	edycja
portal	**baza**	**jakość**	**jakość**	źródło
serwis	**informacja**	dostępny	**serwis**	kryterium
techniczny	rekord	dziedzina	**Internet**	**Internet**
tematyczny	dziedzina	politechnika	**wyszukiwać**	**serwis**
wyszukiwać	dostępny	**techniczny**	online	**jakość**
	portal	źródło	zasób	online
	dostęp	zasób	źródło	**wyszukiwać**
	online	online	**techniczny**	dziedzina

Słowa kluczowe autorskie	Słowa kluczowe generowane automatycznie, sortowane według:			
	frekwencji (średnio 54%)	**liczby dokumentów** (średnio 43%)	**wartości IDF** (średnio 35%)	**wartości tf-idf** (średnio 48%)
	serwis	**wyszukiwać**	politechnika	politechnika
	wyszukiwać	**Internet**	dziedzina	dostępny
	tworzyć	**serwis**	dostępny	**baza**

Źródło: Opracowanie własne na podstawie analizy przykładowych artykułów.

Żadna proponowana miara modyfikowania wagi leksemu nie przyczyniła się do polepszenia wyniku uzyskanego dla listy leksemów posortowanych według wartości frekwencji.

Nierozwiązanym został problem precyzyjnego wskazywania słów mogących charakteryzować treść tekstu. Przy średnim pokryciu leksemów z zestawów autorskich słów kluczowych na poziomie 54% tylko 9 lematów z 17 uwzględnianych było zgodnych z leksemami autorskimi. Nie wykryto natomiast powtarzającego się wzorca, który zapewniłby akceptowalny poziom pewności przy wyborze z listy automatycznie generowanej tylko leksemów zgodnych z treścią. Z podobną trudnością zmierzył się zespół badający możliwość automatycznego generowania profili semantycznych leksemów. Również w tych badaniach okazało się, że znaczenia przypisywane danemu leksemowi na podstawie jego charakterystyk frekwencyjnych nie odzwierciedlają w pełni znaczeń przypisywanych kognitywnie[215].

[215] Zob. A. Pawłowski, M. Piasecki, B. Broda: *Możliwości i ograniczenia metody automatycznego generowania profili semantycznych leksemów na podstawie danych korpusowych. Przykład polskich symboli kolektywnych*. W: *Zeszyty prasoznawcze 2010*, pod red. W. Pisarka, nr 34 (203/204). Kraków: OBP UJ 2010, s. 70-77.

Podsumowanie

Analiza wyników badań przeprowadzonych na potrzeby niniejszej książki pozwoliła zweryfikować wyznaczone cele oraz postawione hipotezy badawcze. W trakcie prac badawczych porównano tradycyjne oraz automatyczne metody tworzenia charakterystyk wyszukiwawczych dokumentów. Ze względu na założoną tematykę pracy, skoncentrowaną głównie na słowach kluczowych jako metainformacji dotyczącej treści, szczegółowej analizie poddano dwa podzbiory słów kluczowych. Jeden z nich stanowiły wyrażenia kluczowe tworzone w wyniku indeksowania tradycyjnego, zaś drugi podzbiór był generowany automatycznie w wyniku operacji komputerowego przetwarzania tekstów języka naturalnego. Badania przeprowadzono na artykułach, których teksty utworzyły dwa niezależne badawcze korpusy tekstów[216].

Dobór materiału pozwolił na wyodrębnienie zestawów słów kluczowych wskazanych przez autorów analizowanych tekstów. Dodatkowo w wyniku badań uzyskano drugi zbiór słów kluczowych pochodzących od indeksatorów (podejście kognitywne). Porównanie odpowiednich zestawów metainformacji przypisanych tym samym tekstom dostarczyło interesujących wniosków.

[216] Jak podano w rozdziale poświęconym prezentacji korpusu badawczego, były to teksty z zakresu informacji naukowej (tworzące główny korpus tekstów) oraz z zakresu ekonomii i zarządzania (zbiór kontrolny).

Jedna z konkluzji dotyczy dysproporcji w liczbie leksemów użytych przez dwie różne kategorie osób analizujących rzeczowo treść artykułów. Autorzy opisując własne teksty stosowali prawie dwa razy mniej jednostek leksykalnych niż osoby postronne (dla przypomnienia: doświadczone w zakresie opracowania rzeczowego dokumentów). Metainformacje przygotowane przez indeksujących studentów były bardziej szczegółowe niż autorskie. Stan ten może mieć przyczynę w fakcie redakcyjnego ograniczenia formalnego, narzuconego autorom przygotowującym teksty do publikacji, określającego dopuszczalną liczbę podanych wyrażeń (lub słów) kluczowych. Studenci – indeksatorzy tworzący charakterystyki treści dokumentów w trakcie badań – nie byli ograniczeni takimi formalnymi ustaleniami. Kolejnym powodem oszczędniejszego wskazywania słów kluczowych przez autorów może być dodatkowa możliwość (wynikająca z wymogu formalnego) utworzenia charakterystyki treści artykułu w postaci abstraktu. Dzięki temu zbiór metainformacji przypisanych jednemu tekstowi mógł zostać rozbity na dwa uzupełniające się podzbiory o różnym charakterze (słowa kluczowe oraz tekst ciągły). Indeksatorzy dysponowali wyłącznie możliwością wskazania słów/wyrażeń kluczowych. Ostatecznie można stwierdzić z dużą dozą prawdopodobieństwa, że słowa kluczowe autorskie uwzględniają najczęściej dodatkowe czynniki pozatreściowe, jak np. profil docelowej publikacji lub grupy odbiorców. Należy jednak zauważyć, że wyższy poziom szczegółowości charakterystyk przygotowanych przez indeksatorów nie zawsze przekłada się na wzrost poziomu efektywności i skuteczności wyszukiwania. Nadmiar wyrażeń kluczowych przypisanych do jednego dokumentu może doprowadzić do wskazywania użytkownikom odpowiedzi nierelewantych (lub relewnatnych w niewielkim stopniu) do ich zapytań.

Kolejny wniosek dotyczy stopnia zgodności słownictwa obu zbiorów wyrażeń kluczowych przygotowanych tradycyjnie. Wbrew oczekiwaniom, średnio ponad dwukrotna przewaga liczebna słownictwa użytego przez indeksatorów nad słownictwem wykorzystanym przez autorów nie zapewniła stuprocentowej zgodności leksyki w obu grupach wyrażeń. Średni stopień zgodności leksykalnej pomiędzy zestawami utworzonymi w wyniku procesów kognitywnych kształtował się na poziomie ok. 75%. Poziom ten wydaje się być górną granicą zgodności indeksowania tradycyjnego przeprowadzanego na tych samych dokumentach przez różne osoby. Badania zgodności indeksowania wykazały, że prawdopodobieństwo wska-

zania tego samego wyrażenia indeksującego (relewantnego dla danego dokumentu) przez różne osoby indeksujące treść dokumentu jest zazwyczaj niższe niż 75%[217].

Natomiast słownictwo o najwyższych pozycjach rankingowych na listach generowanych automatycznie, w zależności od zastosowanej metody określania wagi leksemów, cechowało się zgodnością z autorskimi słowami kluczowymi na poziomie od 54% w przypadku przyjęcia częstości wystąpień jako kryterium do 44% dla współczynnika *idf*. Odpowiednio średni stopień podobieństwa zbiorów leksemów generowanych automatycznie oraz wskazanych przez indeksatorów wynosił od 52% do 49%.

Otrzymane wyniki pozwalają stwierdzić, że postawiony przed niniejszą książką cel, jakim była analiza i ocena możliwości automatycznego generowania słów kluczowych, można zrealizować tylko do pewnego stopnia. Uzyskane poziomy zgodności zbiorów słownictwa, otrzymanego automatycznie i wskazanego w wyniku procesów indeksowania tradycyjnego, są zbyt niskie, żeby uznać proces automatyczny za wystarczający i równoważny z opracowaniem tekstu przez człowieka. Jednakże listy leksemów otrzymane automatycznie, zgodnie z zasadami wykorzystanymi w niniejszej książce, mogą z powodzeniem wesprzeć proces opracowania rzeczowego dokumentów przez człowieka. Dysponując takim narzędziem osoby przygotowujące metadane z charakterystykami treści dokumentów mogłyby skrócić proces opracowania. Komputer znacznie szybciej dokona analizy lingwistycznej tekstu niż człowiek. Połączenie metody automatycznej i tradycyjnej, chociażby w przypadku generowania słów kluczowych opisujących treść, pozwoli znacznie obniżyć koszty całego procesu opracowania rzeczowego[218].

---

[217] Por. Zunde Parnas, M. E. Dexter: *Indexing consistency and quality.* "American Documentation" 1969, vol. 20, nr 3, s. 259-264; C. W. Cleverdon: *Evaluation of operational information retrieval systems.* P. 1: *Identification of criteria*. Cranfield 1964; tegoż: *The Cranfield tests on index language devices*. W: "ASLIB proceedings" 1967, vol. 19, nr 6, s. 173-193; tegoż: *Progress in documentation: evaluation tests on information retrieval systems.* W: „Journal of Documentation" 1970, vol. 26, nr 1, s. 55-67. Cyt za: J. Woźniak: *Kategoryzacja. Studium z teorii...*, s. 22.

[218] O różnych możliwościach zastosowania metod i narzędzi komputerowego przetwarzania języka naturalnego w bibliotekach, ze szczególnym uwzględnieniem kolekcji elektronicznych lub bibliotek cyfrowych, por. m.in. P. Malak, *Możliwości wykorzystania automatycznej analizy wartości informacyjnej dokumentów elektronicznych w tworzeniu kolekcji bibliotecznych*. W: *Biblioteki wobec nowych zadań,* pod red. E. Głowackiej. Toruń: Wydawnictwo UMK 2004, s. 109-110, 127.

Ponadto proponowana metoda pozwala wskazać słowa kluczowe z pełnego tekstu dokumentu. Oczywiście bibliotekarze mają do dyspozycji wsparcie komputerowych systemów bibliotecznych w postaci słowników haseł przedmiotowych lub tezaurusów, ale jest to mimo wszystko słownictwo kontrolowane, przeznaczone przede wszystkim do opisu dokumentów tradycyjnie gromadzonych w bibliotekach. W przypadku dokumentów elektronicznych dostępnych w Internecie czy bibliotekach cyfrowych, przyzwyczajenia wyszukiwawcze użytkowników kierują się raczej w stronę niekontrolowanego wyszukiwania pełnotekstowego. Słownictwo nie podlegające kontroli może być pobierane z pełnego tekstu, co m.in. umożliwia metoda opisana w niniejszej książce.

Przy założeniu możliwości wsparcia komputerowych metod przetwarzania tekstów języka naturalnego przez człowieka, działającego jako instancja nadzorująca i decyzyjna, można stwierdzić, że proces tworzenia charakterystyk wyszukiwawczych dokumentów (z zawężeniem tych charakterystyk do słów kluczowych) poddaje się procesowi automatyzacji. Słownictwo kluczowe wskazywane w wyniku procesów automatycznego przetwarzania tekstów języka naturalnego jest w około 50% zgodne ze słownictwem wybieranym przez człowieka. Nie pozwala to na całkowitą automatyzację procesu tworzenia charakterystyk wyszukiwawczych dokumentów, ale może stanowić doskonałe wsparcie dla pracy człowieka w tym zakresie. Połączenia metod automatycznych z nadzorem ludzkim może wydajnie poszerzyć możliwości opracowania tekstów.

Badania przeprowadzone i opisane w niniejszej książce pozwoliły potwierdzić jedną z postawionych hipotez badawczych. Usunięcie z analizowanych tekstów słownictwa o niskiej wartości informacyjnej i kolejne operacje automatycznego przetwarzania lingwistycznego, jakim poddano dokumenty w trakcie prac badawczych, pozwoliły wygenerować automatycznie listę leksemów, na której najwyższe pozycje zajmuje słownictwo charakterystyczne. Ze słownika terminów charakterystycznych pochodzą zazwyczaj słowa kluczowe opisujące treść dokumentu.

Innym możliwym zastosowaniem takich list jest np. automatyczne przypisanie dokumentu do określonej kategorii lub klasy tematycznej. Przypisanie takie można przeprowadzić niskim kosztem, oznaczając stopień pokrycia słownictwa listy terminów charakterystycznych przez słownictwo najczęstsze w zoptymalizowanej liście leksemów danego dokumentu. Oczywiście należałoby przeprowadzić bardziej szczegółowe ba-

dania w tym kierunku, ale perspektywa wygląda obiecująco. Na korzyść proponowanego rozwiązania można przywołać niskie koszty operacyjne związane głównie z analizą korpusową tekstów z poszczególnych dziedzin badawczych w celu ustalenia list słownictwa charakterystycznego.

Przy okazji można stwierdzić, że w znacznym stopniu potwierdzona została kolejna hipoteza badawcza dotycząca możliwości ustalania wagi słów wyłącznie na podstawie ich frekwencji. Porównanie list słów kluczowych wskazanych w wyniku procesów automatycznych z listami ustalonymi przez człowieka wykazało, że najwyższy stopień zgodności zachodził w przypadku nadawania wagi słownictwu na podstawie frekwencji lokalnych (w obrębie danego tekstu), zaś stosowana powszechnie metoda ważenia *tf-idf* generowała zbiór leksemów o niższym stopniu zgodności. Warto tu przypomnieć, że analizowane teksty oczyszczane były na początku procesu przetwarzania ze słów o niskich wartościach informacyjnych. Dzięki temu leksemy o najwyższych pozycjach rankingowych, ustalanych na podstawie częstości występowania, tworzyły słownik terminów charakterystycznych. Ponadto przeprowadzono dodatkowo analizę wartości wag przyznawanych leksemom z wykorzystaniem we wzorze *td-idf* logarytmów o podstawie 2 oraz o podstawie 10. Uzyskane wyniki zgodności leksemów kluczowych z ustalanymi przez człowieka były na podobnym poziomie w przypadku obu podstaw logarytmów, co potwierdza wnioski S. Robertsona[219]. W obu przypadkach analizowanych podstaw logarytmu wygenerowane zbiory leksemów cechowały się niską zgodnością ze zbiorami słów kluczowych wskazanych w wyniku procesów kognitywnych.

Z kolei wbrew początkowej, intuicyjnej hipotezie roboczej, okazało się, że słowa wyróżnione w tekście przez autorów nie wpływają na automatyczne generowanie słów kluczowych. Zgodność leksemów wyróżnionych z leksemami kluczowymi utrzymująca się na poziomie 33% jest zbyt niska, żeby w jakikolwiek sposób można było za ich pomocą zwiększyć wagę wybranych leksemów z tekstu.

Na podstawie doświadczeń zdobytych podczas prac badawczych opisanych w niniejszej książce można pokusić się o sformułowanie kilku postulatów o charakterze przeważnie technologicznym.

[219] Por. S. Robertson: dz. cyt., s. 2.

Postulaty technologiczne

Na etapie selekcji dokumentów do korpusu badawczego pojawiły się dwa rodzaje czynników, które mogą mieć wpływ na wyniki i jakość komputerowego przetwarzania tekstów języka naturalnego. Były to czynniki związane z brakiem autorskich słów kluczowych oraz związane z formatami zapisu dokumentów.

Standardy metainformacji

W zakresie dokumentów stanowiących główną część materiału badawczego, czyli związanych tematycznie z informacją naukową i bibliologią, problemem był brak autorskich słów kluczowych w większości analizowanych publikacji. Generalnie można zauważyć, że słowa kluczowe częściej pojawiają się w tekstach z zakresu nauk przyrodniczo-matematycznych, są natomiast stosunkowo rzadkim elementem metainformacji w tekstach humanistycznych, np. w „Przeglądzie Bibliotecznym" pojawiają się regularnie dopiero od roku 2007. Jednakże dostrzegalna jest tendencja do opatrywania również tekstów humanistycznych w zestawy słów kluczowych: im nowsze są publikacje, tym częściej można je w nich spotkać.

Ten brak słów kluczowych opisujących artykuły jest o tyle zastanawiający, że teksty z zakresu informacji naukowej i bibliologii dotyczyły często zagadnień wyszukiwania informacji czy też systemów informacyjno-wyszukiwawczych. A przecież słowa kluczowe są jednym ze sposobów usprawnienia wyszukiwania dokumentów w takich systemach. Na szczęście w nowszych publikacjach można zaobserwować wzrost liczby artykułów opisanych w taki sposób. Na wyróżnienie zasługują zarówno „Przegląd Biblioteczny", jak i „Zagadnienia Informacji Naukowej", gdzie autorzy proszeni są o wskazanie słów kluczowych. Również pozytywnie prezentują się pod tym względem materiały konferencyjne, w których coraz częściej słowa kluczowe stają się standardem.

Formaty zapisu dokumentów

Niezależnie od tematyki analizowanych dokumentów oraz szczegółowości dołączonych do nich metainformacji, w wielu przypadkach pojawiał się problem pozyskania tekstu z dostępnych plików.

Ze względu na to, że procesy analizy treści i wskazywania słów kluczowych były realizowane automatycznie, wszystkie badane dokumenty pobierane były w postaci plików .PDF lub .DJVU. Oba formaty są bardzo popularne w cyfrowej publikacji treści, szczególnie w przypadku publikacji w sieci Internet. Format .PDF jest powszechnie wykorzystywany do publikowania dokumentów we własnym zakresie danej jednostki organizacyjnej, zarówno w przypadku zamieszczania własnych artykułów przez konkretnych autorów, jak i w publikacjach dokonywanych przez instytucje. Z kolei format .DJVU jest najpopularniejszym formatem instytucjonalnego publikowania dokumentów. Jest on powszechnie stosowany przez biblioteki wirtualne. Wynika to z jego zalet, wśród których najważniejszą jest korzystny stosunek jakości cyfrowego odwzorowania dokumentu do rozmiaru pliku wynikowego. Stosunek ten jest zdecydowanie korzystniejszy niż w przypadku dokumentów .PDF[220].

Należy wspomnieć, że zauważalnym problemem przy automatycznej analizie treści dokumentów elektronicznych publikowanych w jednym z wymienionych formatów były kłopoty z ekstrakcją tekstu z plików. W przypadku plików .PDF przyczyną takiej sytuacji może być np. nieprawidłowe zaimplementowanie czcionek z dokumentu oryginalnego w trakcie konwersji na format .PDF. W takiej sytuacji utrudnione, a niekiedy nawet niemożliwe jest automatyczne pobranie tekstu z pliku. Można posiłkować się kopiowaniem treści do edytora tekstu i samodzielną korekcją błędów. Jest to jednak proces wysoce czasochłonny. Sytuacji takiej można uniknąć stosując w dokumentach przeznaczonych do konwersji na format .PDF standardowy, systemowy zestaw czcionek.

W przypadku plików .DJVU problemy z pobraniem tekstu mogą być wynikiem technologii skanowania dokumentów – ze względu na oszczędność czasu i miejsca na dyskach strony digitalizowanych dokumentów zapisywane są jako obrazy i konwertowane do standardu .DJVU. W takim przypadku dosyć skutecznym rozwiązaniem okazało się przekonwertowanie pliku do formatu tekstowego. W przypadku takiej konwersji, nawet dla dokumentów zapisanych jako obrazy, informacje pobierane są z warstwy tekstowej. Aplikacją przydatną do pozyskiwania tekstu z plików .DJVU jest m.in. pakiet DjVuLibre, dostępny na stronie: http://djvu.sourceforge.net/.

---

[220] Zarówno bardziej szczegółowy opis zalet i wad obu formatów, jak i analiza stopnia wykorzystania ich w różnych formach udostępniania dokumentów elektronicznych, zagadnienia niewątpliwie bardzo interesujące, wykraczają poza zakres tematyczny niniejszej książki.

Propozycje dalszych badań

Z całości dotychczasowych rozważań wynika, że dokumenty elektroniczne, po zapewnieniu pełnego dostępu do ich treści, można półautomatycznie opisywać za pomocą słów kluczowych. System działający według zasad opisanych w niniejszej książce może wspierać osoby indeksujące treść dokumentów w ich obowiązkach. Pomoc ta wynika chociażby z dużo krótszego czasu pracy systemu (komputera) nad dokumentem w porównaniu z człowiekiem. Ponadto przydatne mogą okazać się wygenerowane automatycznie listy słów kluczowych, z których pracownik informacji może wybrać odpowiedni zestaw tworzący charakterystykę dokumentu.

Zaprezentowane w niniejszej książce metody przetwarzania języka naturalnego w odniesieniu do procesów indeksowania i wyszukiwania informacji pozwalają uzyskać od systemu komputerowego wsparcie w procesie tworzenia charakterystyk treści dokumentów. Oczywiście metody te wymagają dalszych prac nad poprawieniem jakości generowanych wyników. Tę poprawę można osiągnąć za pomocą metod stosowanych w systemach uczących się poprzez, chociażby, wskazanie systemowi, które z proponowanych automatycznie słów kluczowych są relewantne. Innym potencjalnym sposobem podniesienia jakości jest wykonanie dokładniejszych analiz frekwencyjnych słownictwa w zawężeniu do poszczególnych dyscyplin. Szczególnie ten ostatni kierunek wydaje się być obiecujący w kontekście automatycznego przetwarzania języka naturalnego. Zaawansowane charakterystyki frekwencyjne, już nie tylko na poziomie języka czy stylów funkcjonalnych, ale na poziomie tekstów poszczególnych dyscyplin, pozwoliłyby uniknąć sytuacji oznaczenia danego leksemu w funkcji występującej marginalnie (por. oznaczenie wyrazu *do* jako rzeczownika) w procesie wstępnej analizy tekstów. Ponadto można pokusić się o wskazanie przynależności danego tekstu do konkretnej dyscypliny za pomocą szczegółowych wskaźników częstości występowania wyrazów. Dostęp do danych frekwencyjnych jest również przydatny przy badaniach nad automatycznym tłumaczeniem tekstów.

Niewątpliwie system automatycznie generujący słowa kluczowe relewantne do treści dokumentu przyczyniłby się do poprawy wyników wyszukiwania informacji w rozległych repozytoriach cyfrowych. Zaś popularność metody wyszukiwania informacji za pomocą słów kluczowych, obserwowana wśród użytkowników wyszukiwarek internetowych, uzasadnia podejmowanie dalszych badań w tym zakresie.

Bibliografia

1. *11.8.4. Full-Text Stopwords*. W: *MySQL 5.0 Reference Manua*l [on-line]. [Dostęp: 15 października 2011]. Dostępny w World Wide Web: http://dev.mysql.com/doc/refman/5.0/en/fulltext-stopwords.html.
2. Adamczewski P.: *Słownik informatyczny*. Gliwice: Wydawnictwo Helion 2005.
3. *Automatyczne metody konstrukcji sieci semantycznej leksemów polskich na potrzeby przetwarzania języka naturalnego* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://plwordnet.pwr.wroc.pl/main/.
4. Babik W.: *Słowa kluczowe*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego 2010.
5. Białecki A.: *Stempel – algorithmic stemmer for Polish language* [on-line]. Getopt.org [Dostęp: 19 października 2011]. Dostępny w World Wide Web: http://getopt.org/stempel/index.html#distrib.
6. *Biblioteki naukowe w kulturze i cywilizacji. Działania i codzienność. Materiały konferencyjne, Poznań 15-17 czerwca 2005*, pod red. H. Ganińskiej, t.1, 2. Poznań: Biblioteka Główna Politechniki Poznańskiej 2005.
7. *Biblioteki XXI wieku. Czy przetrwamy?: II Konferencja Biblioteki Politechniki Łódzkiej, Łódź, 19-21 czerwca 2006 r. Materiały konferencyjne*. Łódź: Politechnika Łódzka 2006.

8. Bień J. S.: *Aparat pojęciowy wybranych systemów przetwarzania tekstów polskich* [on-line]. Kraków 2006. [Dostęp: 14 listopada 2009]. Dostępny w World Wide Web: http://www.ptj.civ.pl/component/option,com_docman/task,doc_download/gid,20/Itemid,8/.
9. Bień J. S.: *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji* [on-line]. [Dostęp: 14 listopada 2009]. Dostępny w World Wide Web: http://bc.klf.uw.edu.pl/12/2/emph.pdf.
10. Bień J. S.: *O pojęciu wyrazu morfologicznego* [on-line]. Kraków 2006. [Dostęp: 14 listopada 2009]. Dostępny w World Wide Web: http://bc.klf.uw.edu.pl/62/1/jsb-zsE.pdf.
11. Bird S., Klein E., Loper E.: *Natural language processing with Python. Analyzing text with the Natural Language Toolkit*. Sebastopol: O'Reily 2009.
12. Bojar B.: *Językoznawstwo dla studentów informacji naukowej.* Warszawa: Wydawnictwo SBP 2005.
13. Bolc L.: *Natural language parsing systems*. Berlin: Springer Verlag 1987.
14. Daciuk J.: *Narzędzia do automatów skończonych* [on-line]. Gdańsk: Politechnika Gdańska. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://www.eti.pg.gda.pl/katedry/kiw/pracownicy/Jan.Daciuk/personal/fsa_polski.html.
15. *Django. Framework webowy dla perfekcjonistów z terminami* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.django.pl/.
16. *Django. The Web framework for perfectionists with deadlines* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.djangoproject.com/.
17. *Ekonomia XXXIX. Nauki humanistycznospołeczne, Zeszyt specjalny. Dynamiczne modele ekonometryczne, z. 389*, pod red. M. Piłatowskiej. Toruń: Wydawnictwo UMK 2009.
18. *Ekonomia XXXVIII. Nauki humanistycznospołeczne, z. 388*, pod red. M. Piłatowskiej. Toruń: Wydawnictwo UMK 2008.
19. *Electronic Statistics Textbook* [on-line]. Tulsa: StatSoft, Inc. 2001. [Dostęp: 11 stycznia 2012]. Dostępny w *World* Wide Web: http://www.statsoft.com/textbook/.

20. *Encyklopedia językoznawstwa ogólnego*, pod red. K. Polańskiego. Wrocław: Zakład Narodowy im. Ossolińskich 1999.
21. *Encyklopedia PWN* [on-line]. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://encyklopedia.pwn.pl/haslo.php?id=3917948.
22. Fourner J.: *Folksonomies*. W: *Encyclopedia of library and information sciences*, 3rd ed. Boca Raton (FL) 2010.
23. *GB Soft – archiwizacja, udostępnianie dokumentacji. Oprogramowanie DjVu, djvu viewer* [on-line]. [Dostęp: 21 kwietnia 2010]. Dostępny w World Wide Web: http://www.djvu.com.pl/download.php.
24. Gawrysiak P.: *Klasyfikacja. Narzędzie zarządzania i wyszukiwania informacji.* Warszawa: MOST Press 2009.
25. Głowacka E.: *Badania efektywności języków informacyjno-wyszukiwawczych (komunikat z badań).* W: *Komputeryzacja bibliotek. Materiały konferencji 24-26 maja 1993 r.*, pod red. B. Ryszewskiego. Toruń: Wydawnictwo UMK 1994, s. 209-214.
26. Głowacka E.: *Biblioteki wobec nowych zadań*. Toruń: Wydawnictwo UMK 2004.
27. Głowacka T.: *Analiza dokumentu i jego opis przedmiotowy*. Warszawa: Wydawnictwo SBP 2003.
28. Głowacka T.: *Kartoteka wzorcowa języka KABA. Stosowanie w katalogowaniu przedmiotowym*. Warszawa: Wydawnictwo SBP 1997.
29. Gonet K.: *Dlaczego słowa kluczowe a nie hasła przedmiotowe? Co dalej z opracowaniem rzeczowym w bibliotekach FIDES?* W: *FIDES – Biuletyn Bibliotek Kościelnych* [on-line]. 2004, nr 1-2 (18-19), s. 22-32. [Dostęp: 19 kwietnia 2011]. Dostępny w World Wide Web: http://digital.fides.org.pl/dlibra/docmetadata?id=29&from=&dirids=1&ver_id=4698&lp=1&QI=!94F2E9847571EDDCB7D673E25382F1B2-7.
30. Hammerl R., Sambor J.: *O statystycznych prawach językowych*. Warszawa: Zakład Semiotyki Logicznej Uniwersytetu Warszawskiego: Polskie Towarzystwo Semiotyczne 1993.
31. Hammerl R., Sambor J.: *Statystyka dla językoznawców*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego 1990.

32. *Informacja dla nauki a świat zasobów cyfrowych*, pod red. H. Ganińskiej. Poznań: Biblioteka Główna Politechniki Poznańskiej 2008.
33. *i-słownik.pl (słownik slangu informatycznego)* [on-line]. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://www.i-slownik.pl/1,738,jezyki,programowania.html.
34. Jackson P., Moulinier I.: *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. Amsterdam/Philadelphia: John Benjamins Publishing Company 2002.
35. *Językoznawstwo w Polsce: stan i perspektywy*, pod red. S. Gajdy. Opole: PAN – Komitet Językoznawstwa, Uniwersytet Opolski – Instytut Filologii Polskiej 2003.
36. Jurafsky D., Martin J. H.: *Speech and language processing. An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. New Yersey: Prentice Hall 2000.
37. Kamińska-Szmaj I.: *Różnice leksykalne między stylami funkcjonalnymi polszczyzny pisanej. Analiza statystyczna na materiale słownika frekwencyjnego*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego 1990.
38. Kazan A., Skubała E.: *Polskie biblioteki cyfrowe na platformie dLibra – zasób w kontekście tworzenia nowoczesnych kolekcji źródeł informacji dla nauk technicznych*. W: *Informacja dla nauki a świat zasobów cyfrowych*, pod red. H. Ganińskiej. Poznań: Biblioteka Główna Politechniki Poznańskiej 2008, s. 21-34.
39. Kempa A.: *Zastosowanie rozszerzonej metodologii wnioskowania na podstawie przypadków – Textual CBR w pracy z dokumentami tekstowymi* [on-line]. *Systemy wspomagania decyzji. Archiwum publikacji.* Katowice: Akademia Ekonomiczna, Katedra Informatyki 2003-2010. [Dostęp: 19 września 2011]. Dostępny w World Wide Web: http://www.swo.ae.katowice.pl/_pdf/221.pdf.
40. Kłopotek M. A.: *Inteligentne wyszukiwarki internetowe.* Warszawa: Akademicka Oficyna Wydawnicza EXIT 2001.
41. *Korpus języka polskiego IPI PAN* [on-line]. Warszawa: Polska Akademia Nauk, Instytut Podstaw Informatyki 2005-2008. [Dostęp: 22 października 2011]. Dostępny w World Wide Web: http://korpus.pl/.

42. *Korpus języka polskiego Wydawnictwa Naukowego PWN* [on-line]. PWN. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://korpus.pwn.pl/.
43. *Korpus referencyjny języka polskiego PELCRA* [on-line]. [Dostęp: 22 października 2011]. Dostępny w World Wide Web: http://korpus.ia.uni.lodz.pl/index.php.
44. Kupść A., Hajnicz E.: *Przegląd analizatorów morfologicznych dla języka polskiego*. Warszawa: IPI PAN 2001.
45. Kurcz I. i in.: *Słownictwo współczesnego języka polskiego. Listy frekwencyjne*. Warszawa: Instytut Badań Literackich PAN 1974.
46. Kurcz I.: *Pamięć*. W: *Pamięć. Uczenie się. Język*, pod red. T. Tomaszewskiego. Warszawa: Wydawnictwo Naukowe PWN 1995.
47. Kurcz I., Polkowska A.: *Interakcyjne i autonomiczne przetwarzanie informacji językowych. Na przykładzie procesu rozumienia tekstu czytanego na głos*. Wrocław: Zakład Narodowy im. Ossolińskich 1990.
48. Lewandowska-Tomaszczyk B.: *Metody empiryczne i korpusowe w językoznawstwie kognitywnym*. W: *Metodologie językoznawstwa. Podstawy teoretyczne*, pod red. P. Stalmaszczyka. Łódź: Wydawnictwo Uniwersytetu Łódzkiego 2006, s. 262-264.
49. *Lista możliwości wyszukiwania dostępna dla wyszukiwarki Poliqarp dla NKJP* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://nkjp.pl/poliqarp/poliqarp.php?lang=pl.
50. Malak P.: *Metody statystyczne w komputerowym przetwarzaniu języka naturalnego*. „Toruńskie Studia Bibliologiczne" 2011, nr 1 (6) s. 49-62.
51. Malak P.: *Możliwości wykorzystania automatycznej analizy wartości informacyjnej dokumentów elektronicznych w tworzeniu kolekcji bibliotecznych*. W: *Biblioteki wobec nowych zadań*, pod red. E. Głowackiej. Toruń: Wydawnictwo UMK 2004, s. 109-128.
52. Malak P.: *Rozwój badań nad przetwarzaniem języka naturalnego*. „Zagadnienia Informacji Naukowej" 2010, nr 2 (96), s. 21-30.
53. Malak P.: *Słowa kluczowe i tagi jako metody swobodnego oznaczania treści dokumentów w środowisku Nowych Mediów*. W: *Zeszyty Wydziału Humanistycznego VI. Prace Medioznawcze*, pod red. A. Pawłowskiego. Jelenia Góra: Karkonoska Państwowa Szkoła Wyższa w Jeleniej Górze 2011, s. 173-187.

54. Manning Ch. D., Raghavan P. Schütze H.: *An introduction to Information Retrieval* [on-line]. Cambridge: Cambridge University Press 2009. [Dostęp: 19 września 2011]. Dostępny w World Wide Web: http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.
55. Manning Ch. D., Schütze H.: *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press 1999.
56. *Metodologie językoznawstwa. Filozoficzne i empiryczne problemy w analizie języka*, pod red. P. Stalmaszczyka. Łódź: Wydawnictwo Uniwersytetu Łódzkiego 2010.
57. *Metodologie językoznawstwa. Podstawy teoretyczne*, pod red. P. Stalmaszczyka. Łódź: Wydawnictwo Uniwersytetu Łódzkiego 2006.
58. Miłkowski M.: *Morfologik* [on-line]. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://morfologik.blogspot.com/.
59. Mykowiecka A.*: Generacja zdań w języku polskim na podstawie reprezentacji ich semantyki*. Warszawa: IPI PAN 1988.
60. Mykowiecka A.: *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym*. Warszawa: Wydawnictwo PJWSTK 2007.
61. Mykowiecka A.: *Planowanie struktury tekstu przy wykorzystaniu RTS*. Warszawa: IPI PAN 1994.
62. Mykowiecka A.: *Przegląd systemów automatycznej generacji tekstów w języku naturalnym*. Warszawa: IPI PAN 1987.
63. Mykowiecka A.: *Text planning*. Warszawa: IPI PAN 1989.
64. Mykowiecka A.: *Wybrane metody formalnego zapisu składni języka naturalnego*. Warszawa: IPI PAN 1990.
65. *Najpopularniejsze zapytania* [on-line]. [Dostęp: 27 grudnia 2010]. Dostępny w World Wide Web: http://szukaj.wp.pl/najpop.html.
66. *Neurosoft Gram 2.3 Demo* [on-line]. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://gram.neurosoft.pl/.
67. *Neurosoft* [on-line]. [Dostęp: 10 stycznia 2010]. Dostępny w World Wide Web: http://www.neurosoft.pl/?page_name=Produkty_Gram.
68. *Nowy słownik poprawnej polszczyzny PWN*, red. A. Markowski. Warszawa: Wydawnictwo Naukowe PWN 2002.
69. *O projekcie NJKP* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.nkjp.pl/.

70. *Oleander Solutions* [on-line]. [Dostęp: 19 października 2011]. Dostępny w World Wide Web: http://www.oleandersolutions.com/stemming/stemming.html.
71. *Open NLP* [on-line]. [Dostęp: 4 kwietnia 2008]. Dostępny w World Wide Web: http://opennlp.sourceforge.net/.
72. Osiński S., Stefanowski J., Weiss D.: *Lingo: search results clustering algorithm based on singular value decomposition* [on-line]. [Dostęp: 19 września 2011]. Dostępny w World Wide Web: http://www.cs.put.poznan.pl/dweiss/site/publications/download/iipwm-osinski-weiss-stefanowski-2004-lingo.pdf.
73. *Overview Python v2.6.5 documentation* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://docs.python.org/.
74. *Pamięć. Uczenie się. Język*, pod red. T. Tomaszewskiego. Warszawa: Wydawnictwo Naukowe PWN 1995.
75. Pawłowski A.: *Empiryczne i ilościowe metody badań wobec naukowego statusu współczesnego językoznawstwa*. W: *Metodologie językoznawstwa. Filozoficzne i empiryczne problemy w analizie języka*, pod red. P. Stalmaszczyka. Łódź: Wydawnictwo Uniwersytetu Łódzkiego 2010, s. 117-131.
76. Pawłowski A.: *Metody kwantytatywne w sekwencyjnej analizie tekstu*. Warszawa: KLF UW 2001.
77. Pawłowski A., Piasecki M., Broda B.: *Możliwości i ograniczenia metody automatycznego generowania profili semantycznych leksemów na podstawie danych korpusowych. Przykład polskich symboli kolektywnych*. „Zeszyty prasoznawcze" 2010, nr 3/4 (203/204), s. 70-77.
78. Pawłowski A.: *Uwagi na temat korpusu języka polskiego (reprezentatywność, aktualność, nazwa*). W: *Językoznawstwo w Polsce: stan i perspektywy*, pod red. S. Gajdy. Opole: PAN, Uniwersytet Opolski 2003.
79. *Pdf to DjVu GUI main page* (official page) [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.trustfm.net/GeneralTools/SoftwarePdfToDjvuGUI.php.
80. *Pelcra, Korpusy* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://pelcra.ia.uni.lodz.pl/?page_id=3&lang=pl.

81. Pęzik P.: *Przewodnik użytkownika* [on-line]. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://www.ebi.ac.uk/~pezik/korpus//mans/PELCRA_man_pl_19-09-2005.pdf.

82. Piasecki M.: *Automatyczne wydobywanie wiedzy o semantyce języka naturalnego z korpusu tekstów*. W: *Metodologie językoznawstwa. Filozoficzne i empiryczne problemy w analizie języka*, pod red. P. Stalmaszczyka. Łódź: Wydawnictwo Uniwersytetu Łódzkiego 2010, s. 143-182.

83. Piasecki M.: *Cele i zadania lingwistyki informatycznej*. W: *Metodologie językoznawstwa. Współczesne tendencje i kontrowersje*, pod red. P. Stalmaszczyka. Kraków: Lexis 2008, s. 252-290.

84. *Polish stopwords* [on-line]. [Dostęp: 12 marca 2010]. Dostępny w World Wide Web: http://www.ranks.nl/stopwords/polish.html.

85. *Poradnia językowa* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web http://poradnia.pwn.pl/lista.php?id=5606.

86. *Proporcje w pełnej wersji sieciowej korpusu* [on-line]. Korpus języka polskiego Wydawnictwa Naukowego PWN. [Dostęp: 11 stycznia 2011]. Dostępny w World Wide Web: http://korpus.pwn.pl/strukt_full.php.

87. Prywata M.: *Oficjalna strona polskiego zbioru słów dla ispella* [on-line]. Sourceforge.net. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://ispell-pl.sourceforge.net/.

88. Przepiórkowski A., Janus D.: *POLIQUARP 1.0: Some technical aspects of a linguistic search engine for large corpora* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://nlp.ipipan.waw.pl/~adamp/Papers/2006-poliqarp/.

89. Przepiórkowski A.: *Korpus IPI PAN. Wersja wstępna.* Warszawa: IPI PAN 2004.

90. Przepiórkowski A.: *Powierzchniowe przetwarzanie języka polskiego*, Warszawa: Akademicka Oficyna Wydawnicza EXIT 2008.

91. *pyPdf* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://pybrary.net/pyPdf/.

92. *Python License* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.python.org/psf/license/.

93. *Python NLTK (Natural Language Toolkit)* [on-line]. Dostępny w World Wide Web: http://nltk.sourceforge.net.
94. *Python Programming Language Official Web Site* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.python.org/.
95. *Python Success Stories* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.python.org/about/success/.
96. *Python v2.6.5 documentation* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://docs.python.org/index.html.
97. *Quotes about Python* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.python.org/about/quotes/.
98. Robertson S.: *Understanding inverse document frequency: on theoretical arguments for IDF* [on-line]. [Dostęp: 13 listopada 2011]. Dostępny w World Wide Web: http://www.soi.city.ac.uk/~ser/idfpapers/Robertson_idf_JDoc.pdf.
99. Sadowska J., Turowska T.: *Języki informacyjno-wyszukiwawcze. Katalogi rzeczowe*. Warszawa: CUKB SBP 1990.
100. Saloni Z.: *Kategoria rodzaju we współczesnym języku polskim*. W: *Kategorie gramatyczne grup imiennych*, pod red. R. Laskowskiego. Wrocław: Zakład Narodowy im. Ossolińskich 1976, s. 43-78.
101. Salton G., Wong A., Yang C. S.: *A vector space model for automatic indexing* [on-line]. "Communications of the ACM" 1975, vol. 18, nr 11, s. 613-620. [Dostęp: 25 września 2011]. Dostępny w World Wide Web: http://openlib.org/home/krichel/courses/lis618/readings/salton75.pdf.
102. Sambor J.: *Językoznawstwo statystyczne dla pracowników informacji naukowej*. Warszawa: CINTE 1978.
103. *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*, pod red. B. Bojar. Warszawa: Wydawnictwo SBP 2002.
104. *Słownik encyklopedyczny terminologii języków i systemów informacyjno-wyszukiwawczych*, pod red. B. Bojar. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego 1993.

105. *Słownik frekwencyjny polszczyzny współczesnej*, Ida Kurcz i in. Kraków: Instytut Języka Polskiego PAN 1990, T. 1, s. 480; T. 2, s. 481-978.

106. *Słownik SJP.pl – odmiany słów* [on-line]. [Dostęp: 12 października 2010]. Dostępny w World Wide Web: http://www.sjp.pl/slownik/odmiany/.

107. SNOWBALL [on-line]. [Dostęp: 19 października 2011]. Dostępny w World Wide Web: http://snowball.tartarus.org/.

108. Sobczyk M.: *Statystyka. Podstawy teoretyczne przykłady – zadania.* Lublin: Wydawnictwo Uniwersytetu M. Curie-Skłodowskiej 1998.

109. Sobczyk M.: *Statystyka*. Wyd. 3 zm. Warszawa: Wydawnictwo Naukowe PWN 2000.

110. Sosińska-Kalata B.: *Klasyfikacja. Struktury organizacji wiedzy, piśmiennictwa i zasobów informacyjnych*. Warszawa: Wydawnictwo SBP 2002.

111. Sosińska-Kalata B.: *Modele organizacji wiedzy w systemach wyszukiwania informacji o dokumentach*. Warszawa: SBP 1999.

112. Spärck Jones K.: *A statistical interpretation of term specificity and its application in retrieval* [on-line]. [Dostęp: 13 listopada 2011]. Dostępny w World Wide Web: http://www.soi.city.ac.uk/~ser/idfpapers/ksj_orig.pdf.

113. *Statystyki wyszukiwarki Google* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.google.com/insights/search/#geo=PL&date=1%2F2010%2012m&cmpt=q.

114. Świdziński M., Derwojedowa M., Rudolf M.: *Dehomonimizacja i desynkretyzacja w procesie automatycznego przetwarzania wielkich korpusów tekstów polskich* [on-line]. [Dostęp: 10 stycznia 2012]. Dostępny w World Wide Web: http://www.mimuw.edu.pl/polszczyzna/PTJ/b/b58_187-199.pdf.

115. Świdziński M.: *Lingwistyka korpusowa w Polsce – źródła, stan, perspektywy* [on-line]. „LingVaria" 2006, nr 1, s. 23-34. [Dostęp: 19 września 2011]. Dostępny w World Wide Web: http://www2.polonistyka.uj.edu.pl/LingVaria/archiwa/LV_1_2006_pdf/02_swidzinski.pdf.

116. *The Django book* [on-line]. [Dostęp: 21 kwietnia 2010]. Dostępny w World Wide Web: http://www.djangobook.com/.

117. *The IDF page* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.soi.city.ac.uk/~ser/idf.html.

118. *The Porter Stemming Algorithm* [on-line]. Martin's Porter home page. [Dostęp: 15 października 2011]. Dostępny w World Wide Web: http://tartarus.org/~martin/PorterStemmer/.

119. *The Python Wiki* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://wiki.python.org/moin/.

120. Tokarski J.: *Fleksja polska*. Wydanie III z uzupełnieniami. Warszawa: Wydawnictwo Naukowe PWN 2001.

121. *Using Python. Release 2.6.5*, pod red. G. Rossum van, F. L. Drake Jr. W: *Using Python – Python v2.6.5 documentation* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://docs.python.org/using/index.html.

122. Vetulani Z.: *Komunikacja człowieka z maszyną. Komputerowe modelowanie kompetencji językowej*. Warszawa: Akademicka Oficyna Wydawnicza EXIT 2004.

123. Weiss D.: *A survey of freely available polish stemmers and evaluation of their applicability in information retrieval* [on-line]. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://www.cs.put.poznan.pl/dweiss/site/publications/download/ltc_092_weiss_2.pdf.

124. Weiss D.: *Dawid Weiss – Lematyzator dla języka polskiego* [on-line]. Poznań: Politechnika Poznańska, Zakład Inteligentnych Systemów Wspomagania Decyzji 2006. [Dostęp: 19 października 2011]. Dostępny w World Wide Web: http://www.cs.put.poznan.pl/dweiss/xml/projects/lametyzator/index.xml?lang=pl.

125. Weiss D., Stefanowski J.: *Web search results clustering in Polish: experimental evaluation of Carrot* [on-line]. [Dostęp: 19 września 2011]. Dostępny w World Wide Web: http://www.cs.put.poznan.pl/dweiss/site/publications/download/iipwm-dweiss-2003.pdf.

126. *WIEM, Portal Wiedzy* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://portalwiedzy.onet.pl.

127. *Wikipedia, Wolna encyklopedia* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://pl.wikipedia.org/.

128. Woliński M.: *Komputerowa weryfikacja gramatyki Świdzińskiego. Rozprawa doktorska przygotowana pod kierunkiem dr. hab. Janusza S. Bienia, prof. UW* [on-line]. [Niepublikowana praca doktorska], [dostęp: 17 grudnia 2009]. Dostępny w World Wide Web: http://www.ipipan.eu/staff/m.wolinski/publ/mw-phd.pdf.

129. Woźniak J.: *Kategoryzacja. Studium z teorii* języków informacyjno-*wyszukiwawczych*. Warszawa: Wydawnictwo SBP 2000.

130. Woźniak J.: *Tendencje w teorii i praktyce języków informacyjno-wyszukiwawczych*. W: *Opracowanie przedmiotowe – osiągnięcia naukowe i praktyka* [on-line]. Warszawa: Wyższa Szkoła Ekonomiczno-Informatyczna 2004, s. 3-25. [Dostęp: 15 listopada 2011]. Dostępny w World Wide Web: http://www.wsei.pl/biblioteka/materialy/jezykinfo_ex.pdf.

131. Woźniak-Kasperek J.: *Wiedza i język informacyjny w paradygmacie sieciowym*. Warszawa: Wydawnictwo SBP 2011.

132. *Xpdf* [on-line]. [Dostęp: 8 listopada 2011]. Dostępny w World Wide Web: http://www.foolabs.com/xpdf/.

133. *Zarządzanie XXXVII. Nauki humanistyczno-społeczne*, z. 387, pod red. M. Haffera. Toruń: Wydawnictwo UMK 2007.

134. *Zasoby do pobrania*. W: *Korpus języka polskiego* [on-line]. [Dostęp: 19 listopada 2011]. Dostępny w World Wide Web: http://korpus.pl/index.php?page=download.

135. „Zeszyty prasoznawcze" 2010, nr 3/4 (203/204).

Spis tabel

Spis ilustracji

Spis wykresów

Indeks rzeczowy

Piotr Malak jest asystentem w Inst
gii UMK w Toruniu oraz członkiem
nego. Jego zainteresowania badawcze dotyczą zarządzania informacją, wyszukiwania informacji w dokumentach, inżynierii lingwistycznej oraz zarządzania czasem i zadaniami. Entuzjasta i praktyk lifehacking'u oraz efektywnego zarządzania czasem.

Prowadził wykłady gościnne na Uniwersytecie w Ankarze, Hogeschool van Amsterdam w Amsterdamie oraz Uniwersytecie Wileńskim.

Książka prezentuje wyniki badań porównawczych nad skutecznością metod automatycznych i kognitywnych w tworzeniu charakterystyk wyszukiwawczych za pomocą słów kluczowych. Na potrzeby badań wykorzystany został autorski system analizy kwantytatywnej tekstów języka polskiego posługujący się metodami statystycznymi do ustalenia i oceny frekwencji wyrażeń językowych w korpusie tekstów.
We wprowadzeniu teoretycznym czytelnik płynnie przechodzi od teorii badań nad przetwarzaniem języka naturalnego, poprzez problematykę nazewnictwa, omówienie jednostek badania kwantytatywnego tekstów, cech statystycznych jednostek leksykalnych i wybrane sposoby reprezentacji treści dokumentów do metod optymalizacji tekstu na potrzeby automatycznego przetwarzania.

Książka, z założenia obejmująca zaawansowane zagadnienia wyszukiwania informacji, kierowana jest do badaczy i pracowników sektora informacyjnego oraz użytkowników informacji cyfrowej. Świetnie sprawdzi się także jako pomoc podczas zajęć poświęconych wyszukiwaniu informacji oraz w codziennej praktyce dla osób związanych z informacją cyfrową, bibliotekarzy i badaczy procesów przetwarzania informacji.

9 788361 464426

Cena 32,00 zł

ISBN: 978-83-61464-42-6