
Paweł Płaneta

KOMPUTEROWA ANALIZA TEKSTU W Dyskursach Medialnych

WYBRANE TECHNIKI KOMPUTEROWEJ ANALIZY TEKSTÓW MEDIALNYCH

Minęło wiele lat od czasu, kiedy problematykę zastosowania komputerów w analizie zawartości, a także procedury amerykańskiego projektu *General Inquirer*¹ po raz pierwszy zaprezentowano polskiemu czytelnikowi². Opisany projekt, zainicjowany w USA w latach 60. XX w., był zbiorem procedur informatycznych do systematycznego rozpoznawania i klasyfikowania w badanym tekście wyrazów i wyrażeń wedle określonych kategorii. W następnych latach *Badacza uniwersalnego* rozwijano, a wprowadzane udoskonalenia dotyczyły głównie

-
- 1 P.J. Stone, D.C. Dunphy, M.S. Smith, D.M. Ogilvie, *The General Inquirer: A Computer Approach to Content Analysis*, The MIT Press, 1966., a także E. Kelly, P. Stone, *Computer Recognition of English Word Senses*, North-Holland Linguistic Series, 1975.
 - 2 W. Pisarek, *Analiza zawartości prasy*, OBP RSW „Prasa – Książka – Ruch”, Kraków 1983, s. 142-144.

tego, co od początku stanowiło jądro systemu, czyli słownika kategorii. Od lat 90. XX w. niektóre elementy tego projektu wdrażano także w badaniach wykonywanych w Pracowni Analizy Zawartości Ośrodka Badań Prasoznawczych UJ. Opisywane poniżej techniki gromadzenia, eksploatacji, analizy i opracowania danych tekstowych należą w istocie do procedur z zakresu *text mining*. Ich celem jest odkrywanie i wydobywanie informacji z szerokiej gamy dokumentów tekstowych zebranych w określone zbiory. Chodzi m.in. o docieranie do określonych tematów czy konceptów obecnych w danym zestawie przekazów, a także kombinację zmiennych liczbowych z nieustrukturyzowanymi danymi tekstowymi.

Listy frekwencyjne wyrazów oraz zasięgi występowania poszczególnych kategorii komputerowej analizy zawartości mogą umożliwić rekonstrukcję „intensywności” (*gatekeepingu*)³ określonych treści, *listy słów kluczowych* (wraz z kontekstami) oddają ich „wyrazistość”, swoistość, specyficzność (*agenda setting*)⁴, natomiast *wskaźniki korelacji poszczególnych kategorii, analizy skupień* oraz rezultaty *analizy czynnikowej* są instrumentami badania „spójności” (porządku strukturalnego, inaczej ramowania czy *framingu*)⁵ badanego dyskursu.

-
- 3 D.M. White, *The „Gatekeeper”: A Case Study in the Selection of News*, „Journalism Quarterly” 1949, vol. 27, s. 383-390; P. Shoemaker, S. Reese, *Mediating the Message: Theories of Influence on Mass Media Content*, White Plains, Longman, New York 1996; P. Shoemaker, M. Eicholz, E. Kim, B. Wrigley, *Individual and Routines Forces in Gatekeeping*, „Journalism & Mass Communication Quarterly” 2001, vol. 78 (2), s. 233.
 - 4 M. McCombs, D. Shaw, *The Agenda Setting Function of the Mass Media*, „Public Opinion Quarterly” 1972, vol. 36, s. 176-187.
 - 5 R. Entman, *Framing: Toward Clarification of a Fractured Paradigm*, „Journal of Communication” 1993, vol. 43 (4), s. 51-58.

ANALIZA FREKWENCYJNA WYRAZÓW JAKO ETAP WSTĘPNY ANALIZY ZAWARTOŚCI

W początkowej fazie CACA (*computer assisted content analysis*) wykorzystuje się użyteczną – przeniesioną z lingwistyki – technikę analizy korpusów tekstowych, czyli zbiorów tekstów języka dobranych ze względu na zamierzony cel badań. Analiza tego rodzaju umożliwia badanie zarówno wypowiedzi aktualnych, jak i studia prowadzone w ramach tzw. archeologii dyskursu. Utworzenie odpowiednich – ze względu na zakładany cel badań – homogennych zbiorów tekstów daje, z jednej strony, możliwość uogólnienia wyników dotyczących strumieni przekazów danego systemu, a z drugiej – szansę analizy porównawczej wielu systemów tekstów lub ich części. Jako kryterium zestawienia korpusów można ustalić instancje nadawcze, grupy odbiorców, rodzaje i gatunki wypowiedzi itp. Materiałem do analizy ilościowej musi być względnie obszerny korpus, ponieważ to nie wypowiedź jest przedmiotem analizy, lecz wielkości w niej zawarte lub wielkości, które wynikają dopiero z tak a nie inaczej dokonanego zestawienia wielu wypowiedzi⁶.

Analiza leksykalna tekstu polega na wczytywaniu danego tekstu do pamięci komputera w procesie konwersji tekstu z tradycyjnej, drukowanej, „materiałnej” formy do postaci cyfrowej. Robi się to poprzez jego skanowanie i obróbkę za pomocą oprogramowania

6 Np. w projekcie badawczym *Struktura wiadomości zagranicznych w nagłówkach „New York Timesa” w latach 1989-2014* zebrano około 25 tys. nagłówków opublikowanych na łamach amerykańskiego dziennika od stycznia 1989 r. do czerwca 2014 r. Zbiór liczył ponad 230 tys. wyrazów (*tokens – running words*), co stanowiło mniej więcej 19 tys. haseł (*types – distinct words*) Wskaźnik TTR (*type/token ratio*) wynosi 8,39. Natomiast *standardised TTR std.dev.* – 41,83 (przy podstawie standaryzacji wynoszącej 1000 słów). Średnia długość wyrazu – jako elementu składowego nagłówka – wynosi 4,94 znaków. W innym projekcie *Administracja jako element dyskursu o państwie w przemówieniach programowych polskich premierów w latach 1989-2012 r.* na potrzeby badań zgromadzono stenogramy wystąpień programowych premierów III RP z lat 1989-2012, które uporządkowano w korpus tekstowy liczący ok. 5,5 tys. wypowiedzeń (tj. zdań lub ich równoważników) zawierających blisko 90 tys. form wyrazowych (słowoform).

umożliwiającego optyczne rozpoznawanie znaków tekstowych. W przypadku analizy przekazów w formie cyfrowej, np. dostępnych online, gromadzenie materiału badawczego jest łatwiejsze⁷.

Początkowym etapem ilościowej analizy leksykalnej tekstów zebranych na potrzeby określonych badań jest zwykle analiza frekwencyjna. Wysoka liczebność występowania określonych słowotform (np. zbioru o wspólnym polu semantycznym) jest zazwyczaj wskaźnikiem nasycenia badanego korpusu określoną tematyką, a także świadczy o tym, że pewne wyrazy (lub grupy słów) uzyskały w analizowanych tekstach szczególny status, były wielokrotnie powtarzane i – co szczególnie ważne z punktu widzenia medioznawcy – pojawiały się w kluczowych fragmentach tekstu, typowych dla prasy drukowanej, czyli nagłówkach, leadach, śródtytułach itp. Co więcej, wysoka frekwencja określonych jednostek leksykalnych – w szczególności tych słów, które nazywamy sztandarowymi – jest dowodem, że nadawca stara się w mocny sposób pozycjonować w świadomości odbiorcy określoną konfigurację aksjologiczną⁸.

7 Ponieważ teksty dostępne w internecie są zapisywane w różnych formatach, więc w procesie gromadzenia materiału do badań należy odpowiednio wcześniej wykonać niezbędne czynności w celu ujednoczenia formy archiwizowanych przekazów. Proces formatowania tekstu do określonej postaci ma zwykle sprowadzić do porównywalnej postaci różne pod względem technicznym teksty. Chodzi przede wszystkim o ujednoczenie kluczowych elementów organizacji tekstu, np. końców zdań, akapitów itp. Eliminuje się m.in. z zebranych zbiorów tekstowych wszelkie elementy merytorycznie nieistotne: np. znaczniki kodu html, wszelkie znaki sterujące, znaki specjalne jak 'spacje nierozdzielające', 'twarde podziały wiersza'. W efekcie procedura oczyszczania tekstów elektronicznych prowadzi do zgromadzenia zbiorów przekazów o identycznej budowie, które można poddać porównawczej leksykalnej analizie statystycznej.

8 W licznych badaniach tego typu prowadzonych w OBP za wysoko notowane słowoformy uznano takie, które pojawiły się w pierwszej setce listy frekwencyjnej. Oczywiście niemal w każdym dłuższym tekście napisanym w języku polskim pierwsze pozycje listy frekwencyjnej zajmuje zwykle przyimek 'w' lub spójnik 'i'. W przypadku języka angielskiego będą to rodzajniki 'a' i 'the'. Nie ulega też wątpliwości, że w tekstach zredagowanych w obu językach w czołówce listy znajdują się także słowa posiłkowe, takie jak 'się', 'nie', 'że', przyimki 'z', 'na', 'do', a także różne formy czasowników posiłkowych, np. 'być' lub 'mieć'. Takie elementy zostają wyłączone z analizy.

Na przykład rekonstrukcja dyskursu o Polsce na łamach prasy codziennej w USA w latach 1989-2009 polegała na analizie źródeł wypowiedzi dziennikarskich zgromadzonych w bazie *America's Newspapers*, obejmującej miliony tekstów prasowych od lat 70. XX w. Do badań zakwalifikowano próbę tekstów liczącą 1050 artykułów, w których pojawił się wyraz 'Poland' (lub 'Polish') – 50 wypowiedzi dziennikarskich z każdego badanego roku (1989-2009). Celowo dobrane wypowiedzi dziennikarskie uporządkowano w korpus tekstów⁹ liczący 4 900 tys. znaków, ponad 800 tys. form wyrazowych¹⁰ zapisanych w ok. 23 tys. akapitów. Jednostką analizy w tej fazie badań jest zazwyczaj forma wyrazowa (słowoforma), będąca odpowiednikiem tradycyjnego terminu „wyraz”, tj. elementu wyodrębnionego w tekście na podstawie segmentacji graficznej za pomocą odstępów. Intensywność określonych cech tekstu określa się liczbą i zasięgiem procentowym występowania poszczególnych wyrazów w badanym zbiorze¹¹.

W omawianym projekcie sporządzono listy frekwencyjne dla poszczególnych lat 1989-2009, w których wzięto pod uwagę wyłącznie słowoformy pełnoznaczeniowe. Okazało się na przykład, że w 2001 roku ranking najwyżej notowanych wyrazów świadczy

9 W sensie językoznawczym korpusem jest zestaw tekstów języka zebrany w celu zbadania jego systemu lub podsystemu. Zwykle dobiera się teksty do korpusu ze względu na zamierzony cel badań. K. Polański (red.), *Encyklopedia językoznawstwa ogólnego*, Ossolineum, Wrocław 1993, s. 318-319.

10 Forma wyrazowa (słowoforma) jest odpowiednikiem tradycyjnego terminu „wyraz”. Słowoformy są to jednostki wyodrębnione w tekście na podstawie segmentacji graficznej za pomocą odstępów (spacji). Inaczej mówiąc, podstawę wyodrębnienia danej słowoformy stanowi wyłącznie jej postać graficzna, za jedną słowoformę w tym sensie uważa się więc wszelkie wystąpienia jednostek o tej samej postaci zewnętrznej, np. 'człowieka' i 'człowieki' to dwie różne słowoformy jednego hasła 'człowiek'.

11 Najwyżej notowanym słowem na liście frekwencyjnej badanego zbioru tekstów amerykańskich o Polsce był wyraz 'communist', który pojawił się w badanym korpusie blisko 1,8 tys. razy. Innymi charakterystycznymi wyrazami pojawiającymi się na szczycie listy frekwencyjnej były 'Solidarity', 'war', 'pope' oraz 'Jewish' i 'Jews'. Ponadprzeciętne liczebności osiągnęły także takie wyrazy jak 'soviet', 'church' oraz 'German' i 'Germany', natomiast najwyżej notowanym nazwiskiem było nazwisko Lecha Wałęsy.

o licznych odniesieniach do nagłośnionej w tym czasie masakry w Jedwabnem ('Jews', 'Jedwabne', 'Poles', 'Jewish', 'nazi', 'massacre'). A zatem najprostszy pomiar liczebności określonych jednostek w badanym tekście może skutecznie naświetlić dominującą w danym czasie narrację.

Analiza frekwencji wyrazów nie prowadzi wyłącznie do rekonstrukcji dominującej tematyki obecnej w badanych zbiorach tekstów. Na przykład ogólny wniosek z analizy współwystępowania 150 najczęściej używanych wyrazów w wystąpieniach programowych premierów RP w latach 1989-2007 można zamknąć w sformułowaniu: *exposé* jest przede wszystkim wypowiedzią o tym, co „rząd będzie robił z państwem”. Okazało się ponadto, że wystąpienia programowe zawierają dużą liczbę swoistych zwrotów inwokacyjnych oraz form uznanych za wykładniki subiektywności relacji o świecie (egotyczność vs kolektywistyczność tekstu).

Bez względu na to, jaki jest polityczny rodowód premiera, jakie są jego intencje i stosowane szczegółowe techniki perswazji, *exposé* zawsze jest selektywnym obrazem rzeczywistości, a filarami tej selekcji jest zawsze to, co nadawca chciałby, aby odbiorca uznał za ważne, konieczne, słuszne, bliskie i powszechne.

Wyniki pomiarów statystycznych, ujęte w listy frekwencyjne, mogą stanowić podstawę do badań dystrybucji (inaczej konkordancji) wyrazów, istotnych ze statystycznego punktu widzenia. Pomiar ten polega na zestawianiu wszystkich elementów danego typu (wyrazów lub grup wyrazów) występujących w badanym zbiorze tekstów wraz z ich kontekstami. Analiza konkordancji, czyli kontekstów występowania najistotniejszych – z punktu widzenia celu badań – jednostek leksykalnych, stanowi podstawę odkrywania wyrazistych kolokacji określonych wyrazów, czyli inaczej „łączliwości” pewnych słów. Rezultaty analiz konkordancji i kolokacji odkrywają przed badaczem wzory współwystępowania określonych

słowoform razem z innymi, co może być podstawą rekonstrukcji wzajemnych związków, wzorów współwystępowania wyrazów odnoszących się do osób, przedmiotów, cech, czynności, stanów itd.¹²

SŁOWA-KLUCZE W RAMACH ILOŚCIOWEJ ANALIZY LEKSYKALNEJ TEKSTU

Jednym z celów analizy statystycznej słownictwa danego tekstu jest uchwycenie charakterystycznych cech odróżniających go od innych tekstów. Służy temu analiza słów-kluczy, czyli wyrazów, które pojawiają się w określonym zbiorze przekazów wyraźnie częściej, niż w innych tekstach i w tzw. korpusie porównawczym (referencyjnym)¹³. W licznych badaniach prowadzonych w OBP UJ wielokrot-

¹² Np. w badaniach nad dyskursem o Macedonii na łamach prasy światowej w latach 1991-2012 (440 tekstów zawierających ponad 300 tys. form wyrazowych, materiał badawczy z bazy danych światowych mediów (tj. ProQuest Newsstand oraz America's Newspapers) zidentyfikowano liczne odniesienia do instytucji politycznych, a czołowymi kolokacjami wyrazowymi ramach omawianej kategorii były 'government', 'country', 'republic', 'state', 'president' i 'international', ale też – co warto zauważyć – 'border'. Wśród wyrażen kontekstowych ze świata polityki najczęściej pojawiły się 'Republic of Macedonia' (lub 'Former Yugoslav Republic of Macedonia'), ale też 'European Union', a z drugiej strony 'Party of Albanians', i w równym stopniu 'president Kiro Gligorov'. Interesującym wyrażeniem pojawiającym się często w badanych tekstach był ponadto związek 'ethnic Albanian minority', natomiast zauważalna obecność wyrażenia 'international community' wskazuje na umiędzynarodowienie konfliktu w Macedonii. Por. M. Kawka, P. Płaneta, *Dyskursy o Macedonii*, Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków 2012.

¹³ Słowami kluczowymi nie są słowa występujące w danym korpusie najczęściej, czyli te z wysokich pozycji listy frekwencyjnej, lecz takie, których zasięg występowania w danym tekście jest większy, niż można się tego spodziewać, analizując korpus referencyjny. Do najważniejszego elementu analizy słów kluczowych należy właściwy dobór korpusu referencyjnego. Po pierwsze, musi być on nieco większy od korpusu badanego, a po drugie, powinien reprezentować określony typ dyskursu, być próbą tekstów specyficznego rodzaju lub szczególnego stosowania języka (np. języka prasy codziennej). Trudno bowiem utrzymywać, że da się w praktyce zebrać teksty reprezentatywne dla każdego dyskursu lub użycia określonego języka w ogóle. Można wprawdzie

nie wykorzystywano technikę obliczania dla określonych zbiorów tekstów ich kluczowych form wyrazowych. W tym celu korzystano m.in. z informatycznych narzędzi do analizy leksykalnej *Wordsmith Tools*. Jest to zaawansowane oprogramowanie, które wyznacza słowoformy kluczowe badanego tekstu, porównując zasięg występowania każdej formy wyrazowej z listy frekwencyjnej w badanym korpusie z zasięgiem tej formy w korpusie referencyjnym.

Aby określić stopień kluczowości (*keyness*) wybranej słowoformy dla danego tekstu (lub zbioru tekstów), bierze się pod uwagę: 1) frekwencję występowania słowoformy w analizowanym tekście, 2) liczbę pozycji na liście frekwencyjnej tekstu analizowanego, 3) frekwencję słowoformy w większym (referencyjnym) tekście, 4) liczbę pozycji na liście frekwencyjnej korpusu referencyjnego. Na liście słowoform kluczowych pojawiają się zatem takie wyrazy, które mają nadspodziewanie wysoką frekwencję (lub nadspodziewanie niską) w porównaniu z korpusem referencyjnym. Słowoformy pojawiające się w badanym tekście częściej, niż można oczekiwać w porównaniu z korpusem referencyjnym, nazywamy pozytywnymi słowoformami kluczowymi.

Lista słów kluczowych jako efekt porównania list frekwencyjnych badanego zbioru tekstów z listami frekwencyjnymi korpusu referencyjnego umożliwia wnioskowanie na temat tematyki, a nawet gatunku i stylu badanego zbioru przekazów tekstowych. Na przykład zestawienie list frekwencyjnych wyrazów występujących w wyborczych tekstach propagandowych określonych partii politycznych z listą frekwencyjną ogólnego korpusu artykułów prasowych publikowanych w czasie kampanii wyborczej ukazuje podobieństwa i różnice w tekstach kampanii wyborczych poszczególnych stronnictw politycznych. Można

przyjąć, że frekwencyjne słowniki języka polskiego nie tylko prezentują bogactwo jego słownictwa, lecz także informują o częstości używania poszczególnych wyrazów. Jednak nawet najdoskonalsze ogólnojęzykowe słowniki frekwencyjne, po pierwsze, zwykle nie spełniają warunku aktualności, nie zwykle istotnego dla medioznawców, a po drugie, mogą okazać się zbyt ogólnie dla potrzeb specyficznych badań.

również porównywać przekazy propagandowe z dyskursem publicznym i dziennikarskim, jeśli uznamy, że zebrane teksty prasowe są dla niego reprezentatywne.

Aby móc sformułować bardziej wnikliwie wnioski, na kolejnym etapie badań stosuje się komputerową analizę zawartości.

KOMPUTEROWA ANALIZA ZAWARTOŚCI – KLUCZ KATEGORYZACYJNY I KODOWANIE

Podstawą funkcjonowania komputerowej analizy zawartości w prezentowanej propozycji metodologicznej są słowniki zawierające grupy wyrazów. Każdy wyraz hasłowy słownika ma przypisaną wartość (lub wartości), odpowiadającą jakiejś kategorii klucza w koncepcji badań¹⁴. W licznych projektach badawczych OBP korzystano np. ze słownika harwardzkiego¹⁵ (Harvard IV-4), kategorii ze słownika wartości H.D. Lasswella¹⁶ oraz kategorii związanych z tzw. wymiarami semantycznymi C.E. Osgooda¹⁷. Tworzenie klu-

14 Należy jednak podkreślić, że wiele słów należy do słowników różnych kategorii.

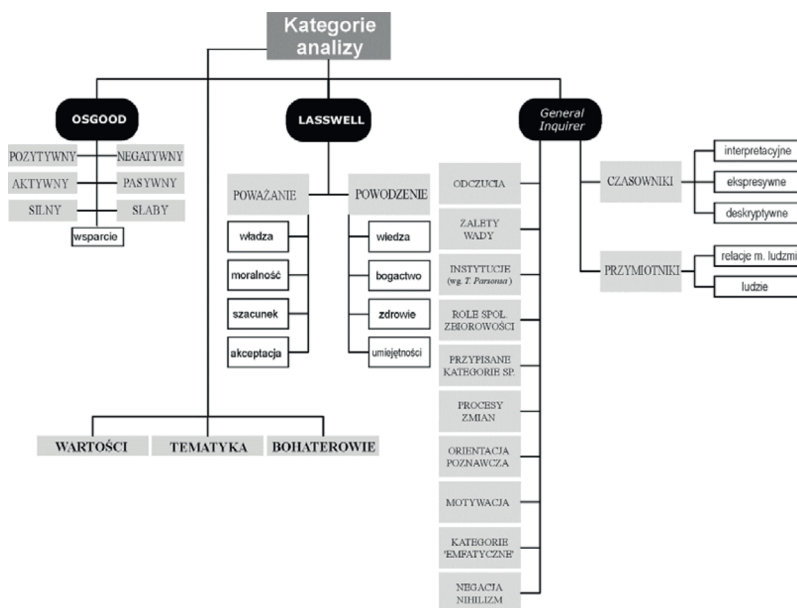
15 W latach 60. w USA stworzono słowniki, które mogły służyć analizie zawartości mediów, a mianowicie *Harwardzki słownik psychosocjologiczny* i *Stanfordzki słownik polityczny*. System kategorii pochodzi z wersji *General Inquirer* zaprojektowanej do analizy zawartości przekazów anglojęzycznych. System ten opracowywano w połowie lat 90. i finansowano ze środków USA National Science Foundation oraz Research Grant Councils of Great Britain and Australia. Autorką najnowszej wersji *GI* napisanej w języku Java jest Vanja Buvac, <http://www.wjh.harvard.edu/~inquirer/> (dostęp: 14.02.2018).

16 Chodzi o wykorzystany w tych badaniach zestaw uniwersalnych kategorii obecnych w słowniku politycznym Harolda D. Lasswella. Pierwszym wymiarem lasswellowskiej analizy dyskursu politycznego jest sfera szeroko rozumianego SZACUNKU, POWAŻANIA, która obejmuje kategorie dotyczące (1) władzy, (2) moralności, (3) prestiżu oraz (4) uczuć. Następny, rozległy wymiar analizy w systemie lasswellowskim, stanowi sfera odnosząca się do POWODZENIA I DOBROBYTU. Wymiar ten obejmuje grupy kategorii odsyłających do (1) wiedzy, (2) bogactwa, (3) pomyślności i (4) umiejętności.

17 Ch.E. Osgood, W.H. May, M.S. Miron, *Cross-Cultural Universals of Affective Meaning*, University of Illinois Press, Urbana Ill. 1975. Opisywana technika

cza można rozpocząć od kategorii uniwersalnych, aby następnie – w zależności od celów konkretnego projektu – konstruować słowniki kategorii szczegółowych, odnoszących się np. do sfery wartości, tematyki przekazów, areny wydarzeń (zwłaszcza regionów, krajów) oraz indywidualnych i kolektywnych bohaterów obecnych w analizowanych przekazach medialnych. Poniżej zaprezentowano ogólny schemat kategorii analizy przekazów tekstowych, zastosowany w licznych projektach badawczych, zrealizowanych w ramach OBP.

Rysunek 1. Schemat ogólnych kategorii analizy zawartości przekazów tekstowych



Źródło: opracowanie własne.

jest znana pod nazwą dyferencjału semantycznego. Uwaga Osgooda koncentrowała się na stronie semantycznej tekstu, autor rozwinął technikę oznaczania różnic między odmiennymi konotacjami słów na kilkustopniowej skali, której przeciwległe bieguny oznaczono przeciwstawnymi cechami.

Podstawą klucza kategoryzacyjnego są słowniki kategorii opracowane z względu na pola wyrazowe, tj. uporządkowane bloki słownika odpowiadające określonym obszarom rzeczywistości¹⁸. Słowniki kategorii zawierają synonimy i wyrazy bliskoznaczne związane z określonymi pojęciami, a kodowanie odbywa się automatycznie, według kwerendy wyszukiwawczej, z wykorzystaniem oprogramowania MS ACCESS. W licznych projektach badawczych realizowanych w OBP, dotyczących zawartości, za jednostkę analizy uznawano wątek tematyczny wypowiedzi, reprezentowany przez akapit tekstu, będący jednostką pośrednią między wyrazem, zdaniem a całą wypowiedzią. Odpowiada on skończonej myśli i jest łatwy do zidentyfikowania i technicznego opracowania¹⁹.

Warto jednak nadmienić, że w licznych projektach OBP jednostkami analizy w procesie CACA były np. nagłówki prasowe albo n-wyrazowe konkordancje istotnych z punktu widzenia badań jednostek leksykalnych. W projekcie *Hate Speech Alert – przeciwko mowie nienawiści w przestrzeni publicznej* jednostką analizy były np. 1000-znakowe otoczenia wyrazów (konkordancje) denotujących poszczególne grupy społeczne.

18 Podstawą zastosowanej w tych badaniach techniki komputerowej analizy zawartości są zbiory wyrazów. Najważniejszą inspiracją konstruowania tychże zbiorów jest koncepcja pola wyrazowego, którą na gruncie polskim spopularyzował wiele lat temu W. Pisarek. W moich badaniach zatem każda kategoria (zbiór wyrazów) to „uporządkowany blok słownika odpowiadający określonemu wycinkowi rzeczywistości percypowanej i analizowanej przez daną społeczność językową”. Por. W. Pisarek, *O mediach i języku*, Universitas, Kraków 2007, s. 278.

19 W sensie technicznym, wszystko to, co wpisano między naciśnięciem klawisza ENTER, jest właśnie akapitem. Programiści stron WWW wprowadzają czasem w miejsce akapitu (ENTER) znak twardego podziału wiersza (tekst ‘od nowej linii w tym samym akapicie’ – SHIFT + ENTER), które na etapie technicznego przygotowania tekstów do analizy skonwertowano do postaci akapitowej.

Rysunek 2. Przykład dystrybucji leksemu 'homoseksualizm' w analizowanym zbiorze tekstów prasowych

N	Concordance	Sign
4 372	właśnie przez prawo do takich działań. Rozmawiał MIROSŁAW SKOWRON Homoseksualizm to nie choroba PROF. ZBIGNIEW LEW-STAROWICZ Seksualność foto	homos*
4 373	walkę z wprowadzeniem wychowania seksualnego do szkół. W styczniu w artykule Homoseksualizm jest zarazliwy w „Rzeczpospolitej” przytoczono wypowiedź	homos*
4 374	walkę z wprowadzeniem wychowania seksualnego do szkół. W styczniu w artykule Homoseksualizm jest zarazliwy w „Rzeczpospolitej” przytoczono wypowiedź	homos*
4 375	w Ciudad del Este. Tam odwołali, zrucając na atakującego go Cuijaję oskarżenie o homoseksualizm . Płapież się nie ugiął i odwołał z urzędu oba skomponowanych	homos*
4 376	wciążęgi ich ich grę, ze przeszytymi się dziwić, etykietował ich postawy. ADHD? Homoseksualizm? Kompleks Edypa? Co z tego? Nie wykłamyj sobie oczu, będziemy	homos*
4 377	biorną się twierdzenia, że to jednostka chorobowa, którą można wyleczyć? - Kiedyś homoseksualizm za chorobę uznawano. Wyreklamowano go jednak z listy chorób.	homos*
4 378	homoseksualizm to nie choroba? Prof. Zbigniew Lew-Starowicz: - Oczywiście, że tak. Homoseksualizm nie jest bowiem żadną chorobą. - Skąd więc, pana zdaniem, biorą się	homos*
4 379	rodziny biologicznej wydymują także potrzebę opisy seksualny. Męski i żeński homoseksualizm oraz pozamateriałskie stosunki płciowe nie będą już postrzegane w	homos*
4 380	rodziny biologicznej wydymują także potrzebę opisy seksualny. Męski i żeński homoseksualizm oraz pozamateriałskie stosunki płciowe nie będą już postrzegane w	homos*
4 381	wulgarna ponomografu, a w Muzeum Narodowym wystawa „Avs Homo Erotica” promuje homoseksualizm (według twórców, pokazuje ona „piękno męski homoseksualnej”).	homos*
4 382	jaki jest interes partiiwa polskiego, interes narodowy, żeby wspierać i reklamować homoseksualizm ? Dziś w Polsce to pytanie retrocenne. - ROZMAWIAMY O POLSCE	homos*
4 383	homoseksualizm to nie choroba? Prof. Zbigniew Lew-Starowicz: - Oczywiście, że tak. Homoseksualizm nie jest bowiem żadną chorobą. - Skąd więc, pana zdaniem, biorą się	homos*
4 384	wulgarna ponomografu, a w Muzeum Narodowym wystawa „Avs Homo Erotica” promuje homoseksualizm (według twórców, pokazuje ona „piękno młodości homoseksualnej”).	homos*
4 385	jaki jest interes partiiwa polskiego, interes narodowy, żeby wspierać i reklamować homoseksualizm ? Dziś w Polsce to pytanie retrocenne. - ROZMAWIAMY O POLSCE	homos*
4 386	popępnąją grzech. Co więcej, istnieją wiele badań naukowych, które pokazują, że homoseksualizm rodzi całe mnóstwo innych patologii i krzywd wyrządzanych innym	homos*
4 387	popępnąją grzech. Co więcej, istnieją wiele badań naukowych, które pokazują, że homoseksualizm rodzi całe mnóstwo innych patologii i krzywd wyrządzanych innym	homos*
4 388	przeciwnikowi łeczy wynika z nieolerancji, bo uznają łecze za symbol promujący homoseksualizm , to też należy łczyć się z ich zdaniem? - Sytuacja tam jest bardziej	homos*
4 389	przeciwnikowi łeczy wynika z nieolerancji, bo uznają łecze za symbol promujący homoseksualizm , to też należy łczyć się z ich zdaniem? - Sytuacja tam jest bardziej	homos*
4 390	w wspomnianej publikacji „Lekcja równości”, że w debacie na temat „Czy homoseksualizm jest grzechem” razem z osobą homoseksualną w żadnym razie nie	homos*
4 391	homoseksualizm w szkołach. W efekcie 50 proc. więcej nastolatków deklaruje homoseksualizm lub nawiązanie stosunków tego typu. Tadeusz Zachurski z	homos*
4 392	uwagę, że za legalizacją homoseksualizmu proszą zmiawiana karpantia promująca homoseksualizm w szkołach. Jej szczegóły i efekty MaszReizistance przedstawiła w	homos*
4 393	w wspomnianej publikacji „Lekcja równości”, że w debacie na temat „Czy homoseksualizm jest grzechem” razem z osobą homoseksualną w żadnym razie nie	homos*
4 394	homoseksualizm w szkołach. W efekcie 50 proc. więcej nastolatków deklaruje homoseksualizm lub nawiązanie stosunków tego typu. Tadeusz Zachurski z	homos*
4 395	przejawem dyskryminacji”. Skargę do KRSiST złożyło Warszawskie Forum Lewiczy. Homoseksualizm czy leśbijstwo to jest choroba. Człowieka szanując, ale choromu	homos*
4 396	rdykalne słowo jak to kardynała Ferdnanda Sebastiana Aguilera z Pampe-łony, który homoseksualizm nazywa morderstwem. Mało tego - młodyści hierarchowie o	homos*
4 397	2014-02-03. KRZYSZTOF KŁOPOTOWSKI, str. A10 Myślę, że homoseksualizm to najwyższe stadium imperializmu lewiczy? Zaprawdę powiadam	homos*
4 398	2014-02-03. KRZYSZTOF KŁOPOTOWSKI, str. A10 Myślę, że homoseksualizm to najwyższe stadium imperializmu lewiczy? Zaprawdę powiadam	homos*
4 399	skali od 0 do 6, o oznacza całkowity brak poczucia do własnej płci, a 6 wyłącznie homoseksualizm . Zboczenia i seks pozamateriałskie są wzięzione u Amerykanów.	homos*
4 400	z 13 listopada ubiegłego roku w Radu Marja, w której dyrektor rozgłosił nazwał homoseksualizm chorobą i powiedział, że chorego trzeba łczyć. Według szefa	homos*
4 401	z 13 listopada ubiegłego roku w Radu Marja, w której dyrektor rozgłosił nazwał homoseksualizm chorobą i powiedział, że chorego trzeba łczyć. Według szefa	homos*
4 402	skali od 0 do 6, o oznacza całkowity brak poczucia do własnej płci, a 6 wyłącznie homoseksualizm . Zboczenia i seks pozamateriałskie są wzięzione u Amerykanów.	homos*

Źródło: opracowanie własne.

Rysunek 3. Fragment jednego z pól konkordancji

06-HOMOSEKSUALIŚCI	
Concordance no 31	
o skrajnie prawicowych poglądach, posła do Izby Gmin z ramienia partii konserwatywnej, kpt. Jacka Maenamary. Razem jeździli w podróże rekonesansowo po nazistowskich Niemczech, które, zgodnie z relacjami Burgessa, sprowadzały się w większości przypadków do homoseksualnych eskapad ze zboczonymi aktywistami Hitlerjugend”. Hitlerjugend – ze względu na orientację pćciową części swych liderów – nazywano w Niemczech potocznie Homojugend, choć np. przywódca tej organizacji Baldur von Schirach był biseksualistą. Z kolei do gejowskich działaczy komunistycznych przyłgnęło określenie Homintern, ukute w latach 30. przez lewicowego poetę i zdeklarowanego homoseksualistę Wystana Hugh Audena. Tworzyli oni swoistą konspirację wewnątrz konspiracji. Burgess uwielbiał penetrować środowiska brunatnych pederastów. Szczególnie bliską znajomość zawarł z Edouardem Pfeifferem – szefem gabinetu francuskiego ministra wojny, a później premiera Edouarda Daladiera. Opowiadał potem, że „wraz z Pfeifferem	

Źródło: opracowanie własne.

Zbiory tekstowe tworzące poszczególne korpusy kodowano i wprowadzano do bazy danych, składającej się z rekordów (przypadków równych przyjętej jednostce analizy) opisanych za pomocą zaprojektowanych do badań pól (zmiennych, czyli kategorii klucza).

Liczba rekordów zapisanych w bazie danych odpowiada liczbie wszystkich akapitów, a tekst każdego akapitu zapisuje się zwykle w jednym z pól typu *memo* rekordu²⁰.

Procedura kodowania przekazów tekstowych przebiega w sposób następujący. Na wstępie konstruuje się listy wyrazów. Każdy zbiór, który można również nazwać słownikiem, tworzy kategorię zbudowaną najczęściej w oparciu o wspólny zakres znaczeniowy lub, w przypadku niektórych kategorii, na podstawie przynależności do tej samej części mowy. Pomiar polega na identyfikacji pojedynczych wyrazów w tekście oraz porównaniu ich ze słowami znajdującymi się w słownikach kategorii badawczych i klasyfikacji rozpoznanych wyrazów do poszczególnych kategorii słownikowych. Przykładem niech będzie słownik wyrazów denotujących bohaterów w kategorii ‘non-adult’²¹:

baby, boy, child, childish, children, girl, grandchild, grandchildren, immature, infant, kid, newborn, playful, playmate, puppy, teenage, teenager, young, youngster, youth, youthful.

W niektórych projektach badawczych podczas identyfikacji wyrazów wykorzystywano procedurę stemmingu, tj. wyodrębnianiu w procesie analizy tekstu rdzenia danego wyrazu (lub wyrazów), a więc części nieodmiennej, np. w kategorii szczegółowej ‘zacofanie’:

²⁰ W każdym rekordzie zawiera się pole typu *memo*, w którym umieszczono jednostkę analizy (akapit tekstu). Pola typu *memo* przyjmują informację w formie tekstowej bez ograniczenia względem liczby znaków, inaczej niż w przypadku pola tekstowego bazy danych, które dopuszcza tekst nieprzekraczający 255 znaków. Ponieważ niektóre akapity znacznie przekraczają tę liczbę, dlatego zastosowano w bazie danych pola *memo* jako miejsca „przechowywania” analizowanych akapitów tekstu.

²¹ Przy konstruowaniu słowników kategorii w badaniach nad tekstami w języku polskim korzystano zwykle ze słownika SJP.PL (wersja: odmiany słów).

```
Like "*" zacof*" Or Like "*" obskur*" Or Like
"* wstecz*" Or Like "*" ciemniact*" Or Like
"* ciemnot*" Or Like "*" ignoranc*" Or Like
"* prymityw*" Or Like "*" kołtuń*" Or Like
"* kołtun*" Or Like "*" ciemnog*" Or Like
"* zaściankow*" Or Like "*" małomiastecz*" Or Like
"* drobnomieszcz*" Or Like "*" prowincjonal*"
Or Like "*" dulszcz*" Or Like "*" reakcyjn*" Or Like
"* średniowiecz*"
```

W następnej fazie lista (kategoria) wyrazów staje się podstawą automatycznej kwerendy kodującej. Wykonuje ona algorytm wyszukiwania i sprawdzania, które spośród jednostek analizy badanego korpusu tekstowego można zakwalifikować do danej kategorii. Każda jednostka analizy (nagłówek, lead, akapit, cała wypowiedź itp.), w której wystąpił przynajmniej jeden wyraz z listy, zostaje zaliczona do danej kategorii.

Skonstruowanie poprawnego, wyczerpującego i jednoznacznego algorytmu kwerendy kodującej należy do jednej z największych trudności badań. Algorytm niekompletny lub błędny może prowadzić do sytuacji, w której kwerenda kodująca może niektóre jednostki analizy pominąć lub błędnie kwalifikować, ponieważ automatyczna procedura – inaczej niż kodujący człowiek – nie jest w stanie samodzielnie dokonywać modyfikacji w trakcie wykonywania zadania. Ponieważ program, w przeciwieństwie do osoby kodującej, nie rozpoznaje np. homonimów czy homogramów, należy analizowane przekazy tagować, tj. oznaczać odpowiednie miejsca w tekście, aby unikać dwuznaczności²².

Jako przykład wykonania procedury niech posłuży kodowanie kategorii 'niedorośli'. W celu automatycznego zaklasyfikowania do tej kategorii wszystkich akapitów, w których pojawiają się

22 W tym celu wykorzystano m.in. składnię opartą na podstawowych operatorach logicznych (AND, OR, NOT).

wyrazy odnoszące się do dzieci i młodzieży, wykonano kwerendę bazy danych skonstruowaną w języku zapytań SQL²³. Wykorzystano w niej także symbole ogólne (tzw. *wildcards*):

```
UPDATE [Kosowo] SET [Kosowo].NIEDOROŚLI = 1.  
WHERE ((([Kosowo].WYPOWIEDŹ) Like "* baby *"  
Or ([Kosowo].WYPOWIEDŹ) Like "* boy *" Or ([Kosowo].  
WYPOWIEDŹ) Like "* child *" Or ([Kosowo].  
WYPOWIEDŹ) Like "* childish *" Or ([Kosowo].  
WYPOWIEDŹ) Like "* children *" Or ([Kosowo].  
WYPOWIEDŹ) Like "* girl *" Or ([Kosowo].WYPOWIEDŹ)  
Like "* grandchild *" Or ([Kosowo].WYPOWIEDŹ)  
Like "* grandchildren *" Or ([Kosowo].WYPOWIEDŹ)  
Like "* immature *" Or ([Kosowo].WYPOWIEDŹ) Like  
"* infant *" Or ([Kosowo].WYPOWIEDŹ) Like "* kid *"  
Or ([Kosowo].WYPOWIEDŹ) Like "* newborn *"  
Or ([Kosowo].WYPOWIEDŹ) Like "* playful *"  
Or ([Kosowo].WYPOWIEDŹ) Like "* playmate *"  
Or ([Kosowo].WYPOWIEDŹ) Like "* puppy *"  
Or ([Kosowo].WYPOWIEDŹ) Like "* teenage *"  
Or ([Kosowo].WYPOWIEDŹ) Like "* teenager *"  
Or ([Kosowo].WYPOWIEDŹ) Like "* young *"  
Or ([Kosowo].WYPOWIEDŹ) Like "* youngster *"  
Or ([Kosowo].WYPOWIEDŹ) Like "* youth *"  
Or ([Kosowo].WYPOWIEDŹ) Like "* youthful *"))
```

Tekst prezentowanej kwerendy można odczytać jako polecenie aktualizacji, czyli wprowadzenia do każdego rekordu bazy danych pt. 'Kosowo' wartości 1 w polu pod nazwą 'NIEDOROŚLI', jeśli w polu 'WYPOWIEDŹ' (pole typu *memo*, gdzie zapisano treść

²³ Język zapytań (z ang. *query language*) – formułuje się w nim zapytania do bazy danych, w odpowiedzi na które uzyskuje się potrzebne zestawienia, zwane też raportami. Do najważniejszych języków zapytań należą standardy języka SQL oraz język zapytań standardu xBASE.

kodowanego akapitu) znajdzie się wartość (tzn. przynajmniej jeden wyraz) między znakami *baby* LUB *boy* LUB *child* LUB ... itd. aż do ostatniego wyrazu słownika *youthful*.

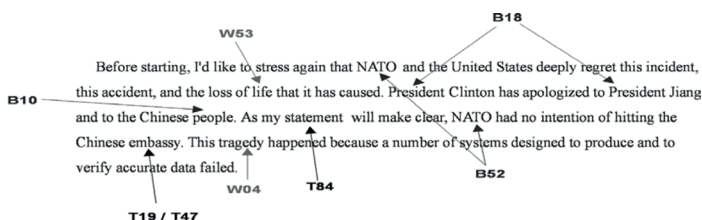
Prezentowany poniżej akapit jest jednym z 400 tys. akapitów zakwalifikowanych do analizy. Został on poddany automatycznej procedurze kodującej w projekcie badań nad propagandą w mediach podczas wojny w Kosowie w 1999 r. W jej wyniku przykładowy tekst akapitu został sklasyfikowany jako odpowiadający następującym kategoriom szczegółowym klucza²⁴: 1) zidentyfikowano następujących BOHATERÓW z listy kategorii: 'ludzie' [kod kategorii – B10], 'politycy' [B18], 'NATO' [B52]; 2) w polu rekordu zawierającym informację na temat KRAJÓW (narodów, narodowości) kwerenda zapisała informację, że w akapicie występują 'USA' i 'Chiny'; 3) rozpoznano nazwiska dwóch AKTORÓW WYDARZEŃ ('Clinton', 'Zemin'); 4) z listy kategorii odnoszących się do pozytywnych i negatywnych WARTOŚCI odnaleziono w tym akapicie 'cierpienie' [W04], 'życie' [W53]²⁵, 'rozwiązywanie problemów'; 5) TEMATYKA podjęta w badanym fragmencie tekstu to 'ambasada chińska' [T19], 'infrastruktura, cele cywilne' [T47] oraz 'dokumenty, oświadczenia, rezolucje itd.' [T84].

Ponadto kwerenda kodująca zidentyfikowała w przykładowym akapicie obecność elementów tekstu, które odpowiadają uniwersalnym kategoriom analizy²⁶.

²⁴ W nawiasach kwadratowych podano zastosowane w omawianym projekcie oznaczenia kodowe poszczególnych kategorii.

²⁵ W rzeczywistości 'loss of life' oznacza 'utrata życia', czyli 'śmierć'. Na tym etapie procedury badawczej postanowiono jednak zachować sztywne kategorie, a uzyskane wyniki ilościowe tych kategorii (np. najliczniej reprezentowane kategorie, wyraźne tendencje do współwystępowania kategorii itd.) poddać szczegółowej analizie w następnej fazie badań.

²⁶ Zidentyfikowano m.in. obecność wszystkich wymiarów Osgooda (tj. wymiar negatywny-pozytywny, słaby-silny, pasywny-aktywny), a także kategorie odnoszące się do ról, zbiorowości i relacji społecznych (np. president), odnaleziono elementy tekstu świadczące o występowaniu form wzmacniających, emfaticznych [emf.01] (I'd like to stress again), ale też osłabiających [emf.01] (np. incident, accident). W prezentowanym fragmencie występują również konstrukcje służące wyrażaniu motywacji działań, np. potrzeby, celu [mot.01,



Po przeprowadzeniu wszystkich procedur kodujących, czyli wykonaniu tylu kwerend, ile zostało zaprojektowanych kategorii klucza, otrzymano bazę danych, w której każdy rekord został oznaczony unikatowym numerem identyfikacyjnym i zawierał informację o jednym akapicie tekstu. Poszczególne pola rekordu – poza polem ‘WYPOWIEDŹ’ – odpowiadały zatem zastosowanym kategoriom analizy, tzn. czy dana jednostka analizy spełniała warunki danej kategorii (wartość pola = 1) czy też nie (wartość pola = 0). Zbudowana w ten sposób baza danych zawierała informacje o korpusie tekstowym zapisane w postaci numerycznej, co umożliwiło operacje statystyczne. Procedura kodująca prowadzi zatem do konwersji strumieni przekazów tekstowych w usystematyzowany zbiór

mot.02] (intention), a także niepowodzenia [mot.07] (failed), zakończenia określonego działania [mot.06] (has caused). Odnajdujemy również słownictwo wskazujące na występowanie bólu [od.02] (regret, apologize), ale brak np. czegokolwiek, co mogłoby świadczyć o obecności elementów przyjemności [od.01]. Również instytucje polityczne (np. embassy, president) były obecne w przedstawianym akapicie, stąd zaliczono ten akapit do kategorii [in.02] z zestawu kategorii *General Inquirer*. Spośród zestawu kategorii lasswellowskich zidentyfikowano kategorię ‘autorytatywni uczestnicy władzy’ [wł.07] (president), ‘etyka’ [LS mo.01] (regret), ‘zdobywanie szacunku’ [LS sz.01] (apologize), ‘zdobywanie wiedzy’ [LS wi.01] (verify), ‘zdrowie – aspekt fizyczny’ [LS zd.03] (*life*, słowo, które zalicza się także do słownika kategorii ‘procesy naturalne’ [zm.01]). Z kolei spośród kategorii dotyczących orientacji poznawczej kwerenda kodująca zidentyfikowała ‘kauzalność’ [ko.03] (caused), ‘relacje’ [ko.10] (systems), ‘rozwiązanie’ [ko.08] (will make clear). W prezentowanym fragmencie odnajdujemy także zaprojektowane do badań kategorie czasowników, tzn. ‘czasowniki-interpretacje czynności’ [CZ.01], ‘służące wyrażaniu uczuć i emocji’ [CZ.02], a także ‘czasowniki opisujące czynności’ [CZ.03], natomiast kwerenda kodująca nie odnalazła przymiotników, przynajmniej takich, które znajdują się w słownikach kategorii ‘przymiotniki opisujące relacje między ludźmi’ [PRZ.01] czy ‘przymiotniki odnoszące się do cech ludzi’ [PRZ.02].

danych liczbowych, poddających się ilościowym opracowaniom statystycznym. Forma zebranych danych i zakodowanych kategorii pozwala np. tworzyć konfiguracje i grupy akapitów spełniających określone kryteria, a także obliczać krzyżowe (lub łączne) wystąpienia kategorii lub ich grup.

„MAPY DYSKURSÓW” JAKO SYNTEZA ANALIZ ZAWARTOŚCI

Aby utworzyć obraz ogólny badanego dyskursu, zebrany materiał badawczy zostaje zwykle poddany różnym procedurom statystycznego wnioskowania wielowymiarowego.

Można np., korzystając z oprogramowania *Text Smart*, skonstruować obraz ogólny na kształt ‘geograficznej’ mapy związków, która bierze pod uwagę zarówno frekwencję pojawiania się określonych wyrazów w poszczególnych jednostkach analizy (nagłówkach), jak i ich współwystępowanie w otoczeniu innych wyrazów. Pierwszy (1) etap opisywanej procedury polega na skonstruowaniu macierzy podobieństwa (*matrix of similarities*) zawartości wszystkich analizowanych przypadków (tj. jednostek analizy, np. nagłówków, akapitów lub n-wyrazowych konkordancji). W tym celu program dobiera w pary wszystkie wyrazy i sprawdza tendencje do współwystępowania każdego wyrazu z pozostałymi²⁷. Dla konstruowania

²⁷ Tzn. jak często każda para wyrazów współwystępuje w poszczególnych nagłówkach, konstruując tabelę czteropolową (tabelę kontyngencji – *2x2 contingency table*) dla każdej pary wyrazów (litery a, b, c, d reprezentują łączną liczbę jednostek analizy (akapitów) spełniającą opisane w nagłówkach rubryk warunki):

		Kategoria 1	
		występuje	nie występuje
Kategoria 2	występuje	a	b
	nie występuje	c	d

Zebrane w ten sposób informacje wykorzystano do obliczenia (dla każdej

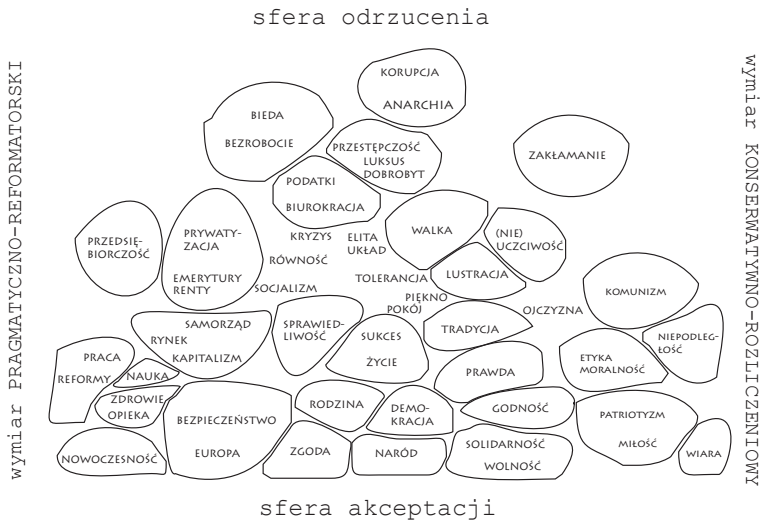
map kategorii (lub podstawowej miary automatycznej kategoryzacji tekstu²⁸) w module *Text Smart* wybrano – spośród wielu typów miar binarnych – współczynnik Jaccarda, ze względu na jego dużą obrazowość²⁹. Kolejny etap (2) prowadzący do utworzenia mapy kategorii polega na hierarchicznym grupowaniu badanych obiektów (na podstawie danych z macierzy podobieństwa) w określoną liczbę skupień. Do wygenerowania określonej liczby skupień *Text Smart* wykorzystuje algorytm będący wariantem hierarchicznej metody grupowania, opartej na tzw. mierze odległości (*hierarchical clustering with maximum distance amalgamation*). Chodzi o to, aby wygenerować skupienia ‘kompaktowe’, tzn. takie, w których najdalsza odległość między jakimikolwiek dwoma ‘członkami’ tego samego skupienia będzie możliwie mała. Efekt finalny (3) opisywanej procedury stanowi przedstawienie utworzonych skupień w układzie graficznym. W tym celu program przeprowadza skalowania wielowymiarowe (*multidimensional scaling*) i obrazuje jego wynik na dwuwymiarowej mapie. Skalowanie wielowymiarowe należy traktować jako środek ukazania pewnych wzorów występowania badanych kategorii, a prezentacja danych w tej formie oczywiście nie jest doskonała. Ponieważ macierz podobieństwa może zawierać (przynajmniej potencjalnie) dużą liczbę wymiarów, ich prezentacja na płaszczyźnie dwuwymiarowej prowadzi do pewnych zniekształceń. Opisana powyżej trzystopniowa procedura doprowadziła

pary wyrazów) miary podobieństwa Jaccarda, którą wyznacza się w następujący sposób – liczbę wspólnych wystąpień dwóch wyrazów (a) dzielimy przez sumę ich wspólnych wystąpień (a) dodaną do sumy wystąpień każdego wyrazu oddzielnie (b i c): $a/(a+b+c)$. Uwaga: z analizy wyłączono słowa posilkowe, spójniki, zaimki itd.

- 28 Należy wspomnieć, że *Text Smart* został zaprojektowany przede wszystkim jako narzędzie do analizy tekstu, np. automatycznej kategoryzacji odpowiedzi udzielanych w pytaniach otwartych.
- 29 Miara Jaccarda jest w pewnym sensie analogiczna do zwykłej miary odległości, całość obliczeń tworzy potężną matrycę podobieństw, która obrazuje „dystans” między parą kategorii, podobnie jak podziałka mapy geograficznej pozwala ustalić odległość między dwoma dowolnymi miastami. Procedurę skalowania wielowymiarowego opisano w *Text Smart 1.0. User's Guide*, SPSS Inc.

do skonstruowania mapy, ilustrującej ogólny porządek dyskursu, uformowanego z wyrazów występujących w wystąpieniach programowych premierów Polski w latach 1989-2007.

Rysunek 4. Wymiar aksjologiczny wystąpień programowych premierów RP w latach 1989-2007

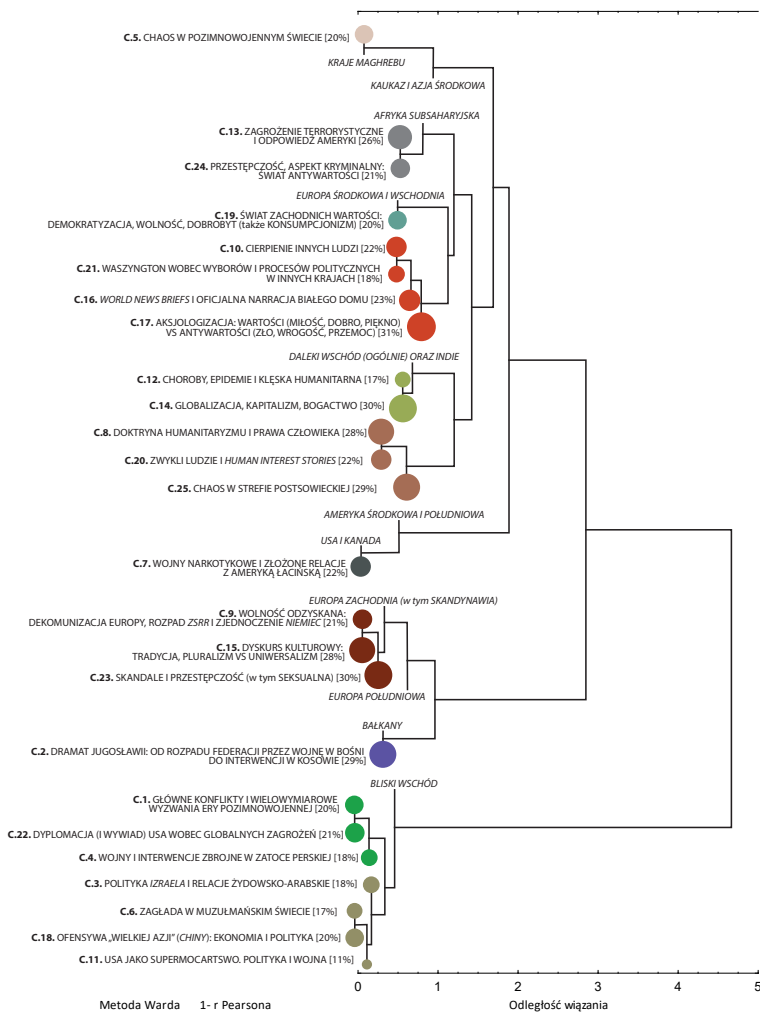


Źródło: opracowanie własne.

Przedstawiony powyżej schemat należy traktować jak mapę, która zawiera kilka obszarów-skupisk wyrazów. Wielkość poszczególnych obszarów na mapie informuje o tym, jak licznie w badanym korpusie dany związek wyrazów był reprezentowany. Z kolei odległości między poszczególnymi obszarami (ale także odległości dzielące poszczególne elementy tych obszarów, tzn. wyrazy) informują o tendencjach do ich współwystępowania w badanych jednostkach analizy, czyli nagłówkach.

Kolejnym sposobem dotarcia do wyrazistej struktury określonych dyskursów jest skorzystanie z pomiarów statystycznych, opartych na technikach wnioskowania wielowymiarowego, zwłaszcza analizy czynnikowej. Procedurę tę zastosowano np. w celu naszkicowania struktury wiadomości zagranicznych w nagłówkach „New York Timesa” w latach 1989-2014.

Rysunek 5. Struktura głównych dyskursów o świecie na łamach „New York Timesa” w latach 1989-2014



Źródło: opracowanie własne.

W wyniku analizy czynnikowej zredukowano obraz całości (25 tys. nagłówków) do 25 najważniejszych czynników (C01-C25), które decydują o tym, że różne, szczegółowe kategorie analizy

są ze sobą mocno związane, mają tendencję do współwystępowania i tworzą tym samym łatwo rozpoznawalną w badanych tekstach konfigurację³⁰.

Następnie, korzystając z techniki statystycznej metody grupowania danych (*cluster analysis*), naszkicowano strukturę współwystępowania kategorii w zbiorze tekstów opublikowanych w rubryce spraw zagranicznych na łamach amerykańskiego dziennika. W tym celu wybrano metodę łączenia skupień (*joining cluster analysis*), prezentowanych w formie rozgałęzień (*tree clustering*) i w oparciu o metodę Warda (pełnego łączenia obiektów – *complete linkage rule of amalgamation*), w której – na podstawie matrycy wskaźników korelacji między 25 czynnikami (oraz kategoriami denotującymi kraje i regiony świata)³¹ – zmierzono odległości między elementami. Wprowadzono ponadto dodatkową informację: wielkość każdego elementu (średnica koła) odpowiada na poniższym schemacie łączonym zasięgom procentowym kategorii tworzących wyodrębnione obiekty.

PODSUMOWANIE

Technologie informatyczne dostarczają dziś skutecznych narzędzi analizy zawartości przekazów tekstowych, stanowiących ważny

³⁰ O tym, czy dana kategoria znalazła się w zestawie określonym przez czynnik X, decydowało jej nasycenie danym czynnikiem (ładunek czynnikowy). W efekcie pewne kategorie mogły wystąpić w kilku zestawach jednocześnie, co wprawdzie zmniejsza dystynktywność czynników, lecz z drugiej strony pokazuje wielostronność, wielowymiarowość określonych kategorii. Innymi słowy, o kategoriach mocno nasyconych ładunkami różnych czynników jednocześnie można powiedzieć, że są bardziej dyskursywne od innych, czyli wyraźniej obecne w przestrzeni badanych przekazów.

³¹ Współczynniki korelacji skonwertowano według wzoru $1 - r$, gdzie r oznacza moment produktowy Pearsona. Np. jeśli współczynnik korelacji między dwiema kategoriami wynosi 0,3, to odległość między nimi $1 - 0,3 = 0,7$. W przypadku korelacji ujemnej $1 - (-0,3) = 1,3$, dlatego na wykresach wartość na osi X (miara odległości) może być większa od 1.

element dyskursów publicznych, w tym medialnych. W proponowanej procedurze można wyodrębnić kilka faz analizy, które – przy zastosowaniu różnorodnych technik pomiaru – mogą wspierać proces wnioskowania.

Listy frekwencyjne wyrazów oraz zasięgi procentowe poszczególnych kategorii komputerowej analizy zawartości mogą służyć rekonstrukcji intensywności określonych treści. Zwykle stanowi to próbę odtworzenia mechanizmów ich selekcjonowania w mediach. Informacje o frekwencji i zasięgach kategorii, połączone ze słowami kluczowymi (wraz z kontekstami), mogą eksponować wyrazistość (swoistość) określonych treści, co pozwala na odtworzenie hierarchii ważności wydarzeń, osób, stanów, czynności (itp.) prezentowanych w mediach. Wykonywana zazwyczaj na ostatnim etapie badań matryca wzajemnych korelacji poszczególnych kategorii badawczych oraz sporządzana na jej podstawie analiza skupień może prowadzić do ujawnienia dominujących ram interpretacyjnych elementów świata przedstawionego, zaś analiza czynnikowa bywa skutecznym narzędziem rekonstrukcji kardynalnych narracji obecnych w medialnych obrazach świata.

BIBLIOGRAFIA

- Berelson B., *Content analysis in communication research*, Free Press, New York 1952.
- Bulandra A., Kościółek J., Zimnoch M., *Mowa nienawiści w przestrzeni publicznej. Raport z badań prasy w 2014 roku*, <http://www.interkulturalni.pl/Elektroniczna-wersja-publicacji--Mowa-nienawisci-w-przestrzeni-publicznej.-Raport-z-badan-prasy-w-2014-roku--juz-do-pobrania-315.html>
- Kawka M., Płaneta P., *Dyskursy o Macedonii*, Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków 2013.

- Kelly E., Stone P., *Computer Recognition of English Word Senses*, North-Holland Linguistic Series, 1975.
- Krippendorff K., *Content analysis: An Introduction to its Methodology*, Sage Publications, Beverly Hills CA 1980.
- McCombs M., Shaw D., *The Agenda Setting Function of the Mass Media*, „Public Opinion Quarterly” 1972, vol. 36, s. 176-187.
- Osgood Ch.E., May W.H., Miron M.S., *Cross-Cultural Universals of Affective Meaning*, University of Illinois Press, Urbana Ill., 1975.
- Pisarek W., *Analiza zawartości prasy*, OBP RSW „Prasa – Książka – Ruch”, Kraków 1983.
- Pisarek W., *O mediach i języku*, Universitas, Kraków 2007.
- Płaneta P., *Obraz Macedonii na łamach polskiej prasy w latach 2000-2007*, [w:] M. Kawka, I. Stawowy-Kawka (red.), *Tożsamość narodowa w społeczeństwie multietnicznym Macedonii*, Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków 2008.
- Płaneta P., *Struktura wiadomości zagranicznych w nagłówkach „New York Timesa” w latach 1989-2014*, „Rocznik Prasoznawczy” 2017, t. 11, s. 131-155.
- Płaneta P., *Słowa sztandarowe w exposé polskich premierów*, „Zeszyty Prasoznawcze” 2009, nr 1-2.
- Płaneta P., *Dwie dekady wolności. Obraz polskiej transformacji na łamach amerykańskiej prasy 1989-2009*, „Zeszyty Prasoznawcze” 2011, nr 2-4.
- Shoemaker P., Reese S., *Mediating the Message: Theories of Influence on Mass Media Content*, White Plains, Longman, New York 1996.
- Stone P.J., Dunphy D.C., Smith M.S., Ogilvie D.M., *The General Inquirer: A Computer Approach to Content Analysis*, The MIT Press, 1966.