UR JOURNAL OF HUMANITIES AND SOCIAL SCIENCES
NR 4(25)/2022 ISSN 2543-8379
ARTYKUŁY DOI: 10.15584/johass.2022.4.7

Agnieszka Hess[1]
Krzysztof Hwaszcz[2]

Językoznawstwo korpusowe w badaniach medioznawczych – ujęcie historyczne i praktyczne

Streszczenie

Celem artykułu jest przedstawienie korzyści i zagrożeń wynikających z implementacji komputerowego językoznawstwa korpusowego do analizy dyskursu. Autorzy opisują genezę i rozwój narzędzi do przetwarzaniu języka naturalnego (z ang. Natural Language Processing, NLP) w ujęciu historycznym oraz prezentują przykłady ich zastosowania w obszarze nauk społecznych, w szczególności w metodologii nauk o komunikacji społecznej i mediach. Praktyczne ujęcie tematu obrazują fragmentaryczne wyniki badań zrealizowanych w Instytucie Dziennikarstwa, Mediów i Komunikacji Społecznej Uniwersytetu Jagiellońskiego we współpracy z konsorcjum CLARIN-PL. Artykuł prezentuje zastosowanie narzędzi NLP w analizie korpusu dyskursu parlamentarnego z lat 1989–2019 pod kątem uwarunkowań instytucjonalizacji dialogu obywatelskiego w Polsce oraz w analizie porównawczej tematu wielokulturowości w dyskursie rady miasta i dyskursie mediów w Krakowie w okresie 2014–2018 (VII kadencja Rady Miasta Krakowa). Autorzy wskazują, w której fazie i jak lingwistyka komputerowa wpisuje się w szeroki kontekst problematyki związanej z badaniami komunikologicznymi – przede wszystkim jako narzędzie, które może wspierać proces wnioskowania.

Słowa kluczowe: analiza dyskursu, analiza mediów, językoznawstwo korpusowe, narzędzia do przetwarzania języka naturalnego

[1] Dr hab. Agnieszka Hess, prof. UJ, Instytut Dziennikarstwa, Mediów i Komunikacji Społecznej, Wydział Zarządzania i Komunikacji Społecznej, Uniwersytet Jagielloński, ul. prof. Stanisława Łojasiewicza 4, 30-384 Kraków, e-mail: agnieszka.hess@uj.edu.pl, nr ORCiD: 0000-0002-4799-9216.

[2] Dr Krzysztof Hwaszcz, Instytut Filologii Angielskiej, Uniwersytet Wrocławski, ul. Kuźnicza 21-22, 50-138 Wrocław, Katedra Sztucznej Inteligencji, Wydział Informatyki i Telekomunikacji, Politechnika Wrocławska, Wybrzeże Wyspiańskiego 27, 50-370, Wrocław, e-mail: krzysztof.hwaszcz@uwr.edu.pl, nr ORCiD; 0000-0003-2136-5001.

Wprowadzenie

Co najmniej od początku lat 90. XX wieku w polskich badaniach medioznawczych stosowane są techniki gromadzenia, eksploatacji, analizy i opracowywania danych tekstowych (Płaneta 2018). Szybki rozwój oprogramowania i komputerowych narzędzi do analizy tekstu, który nastąpił w kolejnych dekadach, umożliwił badaczom magazynowanie danych i zautomatyzowaną obróbkę dużych zasobów językowych.

W komputerowej analizie zawartości, która jest podstawową metodą badawczą w naukach o komunikacji społecznej i mediach, stosowane są coraz częściej – przeniesione z lingwistyki – techniki językoznawstwa korpusowego. Wykorzystuje się je przede wszystkim do badań frekwencyjnych, dzięki którym identyfikowane są spójne semantycznie zbiory słów i wyrażeń wielowyrazowych o najwyższej częstotliwości występowania w materiale badawczym. Słowo (wyraz) stanowi podstawową jednostkę analizowanych danych w badaniach korpusowych. Jego waga jest wypadkową wielu czynników, wśród których podstawowe znaczenie ma właśnie frekwencja użycia (w języku w ogóle, jak również w konkretnym zbiorze wypowiedzi).

Zgodnie z perspektywą przyjętą przez Pisarka (2002) słowa i wyrażenia sztandarowe – zwane także „flagowymi" – zaliczane są do odmiany słów kluczowych, charakteryzujących się trzema istotnymi właściwościami, szczególnie w kwestii badań społecznych: (1) posiadają wysoki wskaźnik wydźwięku emocjonalnego (są nacechowane emocjami, zarówno pozytywnymi, jak i negatywnymi); (2) mają charakter użytkowy (są cenne dla kogoś w danej sytuacji społecznej) i (3) występują w ścisłym związku z systemem wartości wspólnych dla przedstawicieli danej kultury. Metodę badania słów i wyrażeń sztandarowych można zakwalifikować do teoretycznej tradycji badania kluczowości zjawisk zarówno w odniesieniu do kultury, jak i języka (Pisarek 2002). W efekcie sztandarowość uznawana jest za wskaźnik stanu świadomości społecznej.

Analiza frekwencji wyrazów i wyrażeń wielowyrazowych używana jest w medioznawstwie przede wszystkim w celu zidentyfikowania i opisania dominującej problematyki w badanych korpusach. Jej zastosowanie w zaawansowanych formach – przy użyciu narzędzi do przetwarzania NLP – pozwala także na wnioskowanie dotyczące związków i wzorów współwystępowania słownictwa odnoszącego się do osób, instytucji, czynności, stanów czy zjawisk (Płaneta 2018). Co więcej, dzięki wyekscerpowaniu charakterystycznych cech odróżniających jeden zasób tekstów od innych zasobów możliwe jest porównywanie różnych rodzajów dyskursów.

Językoznawstwo korpusowe – początek i rozwój

Początek komputerowego językoznawstwa korpusowego datuje się na lata 50. ubiegłego wieku, ale gwałtowny rozwój nastąpił w latach 70. i 80. (Lewandowska-Tomaszczyk 2005). Wyłonienie się tej dyscypliny językoznawstwa było reakcją na postulaty gramatyki generatywnej Noama Chomsky'ego (1965), w których poczesne miejsce zajmowała intuicja rodzimego użytkownika danego języka. Rola intuicji w stawianiu hipotez dotyczących akceptowalności zdań i ich gramatyczności, jak i w dobieraniu odpowiednich metod badań i analizy jest uważana przez wielu językoznawców za kluczową. Chomsky pojmował jednak tę intuicję jako pewną wyidealizowaną zdolność użytkowników wykorzystywaną przy decydowaniu o gramatyczności zdań poza kontekstem. Model języka w perspektywie gramatyki generatywnej był ujmowany w terminach dychotomicznych, uznających zdanie za albo całkowicie gramatyczne, albo całkowicie niegramatyczne. Z kolei William Labov (1973) zaproponował podejście do gramatyki opierające się na prawdopodobieństwie, że pewne konstrukcje gramatyczne są bardziej preferowane od innych (o czym świadczy ich wysoka lub niska częstość użycia). Częstość użycia poszczególnych konstrukcji gramatycznych w języku nie jest jednak wartością bezwzględną. Jedne struktury mogą być wykorzystywane częściej niż inne w danym dyskursie czy stylu, a wybór form może być zależny od wielu czynników, np. od indywidualnych preferencji użytkownika. Warto tutaj również wspomnieć o semantyce prototypu: częstotliwość użycia nie dotyczy wyłącznie prototypowości na poziomie formy, lecz również wchodzi w strukturę znaczeniową (stereotypy, walencje semantyczne, frazeologia itp.). Te treści semantyczne mają swój wyraz w postaci materialnej: gramatyczno-leksykalnej (Kleiber 1990).

Scalenie perspektyw opierających się na częstości użycia i prawdopodobieństwie występowania danych struktur gramatycznych w różnych odmianach języka doprowadziło do przekierowania uwagi lingwistów na funkcjonalne modele użycia języka. Niezależnie od tego na niespotykaną przedtem skalę rozwinęły się nowe techniki informacyjne, w szczególności komputery i oprogramowanie komputerowe, które umożliwiły badaczom magazynowanie i – wtedy jeszcze tylko częściowo zautomatyzowaną – obróbkę dużych zasobów językowych (Ogrodniczuk 2017; Piasecki 2008).

Jako jedna z gałęzi językoznawstwa komputerowego rozwinęło się językoznawstwo korpusowe, które bada język zgromadzony w korpusach językowych, czyli zdigitalizowanych zbiorach autentycznych tekstów. Pierwszym komputerowym korpusem był stworzony w 1967 r.

przez Kučerę i Francisa korpus amerykańskiej odmiany języka angielskiego, zawierający milion tekstów. Obecne zdigitalizowane zbiory danych sięgają setek milionów jednostek (Lewandowska-Tomaszczyk 2005). Jednym z nich jest Narodowy Korpus Języka Polskiego (NKJP 2021).

Wśród analizowanych korpusów językowych można wyróżnić korpusy zrównoważone, w których zapewnia się odpowiednie proporcje między różnymi stylami tekstów w taki sposób, aby wybrane próbki najtrafniej odzwierciedlały dany język w rzeczywistym użyciu. Innym rodzajem są korpusy celowe, skupiające określony rodzaj tekstów. Należą do nich Korpus Dyskursu Parlamentarnego, który jest olbrzymim, ciągle uzupełnianym zbiorem anotowanych lingwistycznie tekstów z posiedzeń plenarnych Sejmu i Senatu RP, interpelacji i zapytań poselskich oraz posiedzeń komisji z okresu od 1919 r.[3], a także Korpus Dyskursu Rady Miasta Krakowa i korpus dyskursu wybranych mediów z okresu 2014–2018. Zostały przygotowane na potrzeby prezentowanych w niniejszym artykule badań Obserwatorium Dialogu Obywatelskiego (https://dialogobywatelski.org/)[4].

Zastosowanie metod korpusowych przy wykorzystaniu odpowiednich narzędzi oraz cyfrowych baz danych umożliwia znaczne poszerzenie zakresu badań, wyeliminowanie czasochłonnego procesu ręcznej anotacji, prowadzenia manualnych statystyk itp. Ponadto pozwala na weryfikację intuicji językowych, kwantyfikację studiów socjolingwistycznych i dialektologicznych na dużą skalę czy też na badania kontrastywne języka w odniesieniu do różnych grup użytkowników. Przez wzgląd na prezentację i opracowywanie danych w formie elektronicznej językoznawstwo korpusowe wypracowuje własne techniki, metody i narzędzia do analizy tekstu. Z jednej strony metodologia skierowana jest na dochodzenie do uogólnień poprzez sądy indukcyjne, z drugiej zaś

[3] Korpus Dyskursu Parlamentarnego powstał w Zespole Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN. Pomysłodawcą jego stworzenia był prof. Maciej Ogrodniczuk, który nadal koordynuje prace nad aktualizacją i modernizacją zbioru (Ogrodniczuk 2018). Podstawową jednostką analizowanych danych KDP jest słowo, definiowane jako pozycja ze słownika. Sekwencja N słów tworzy dokument, a zbiór dokumentów stanowi korpus.

[4] Obserwatorium Dialogu Obywatelskiego jest projektem badawczo-dydaktycznym realizowanym od 2015 r. przez Wydział Zarządzania i Komunikacji Społecznej UJ (Instytut Dziennikarstwa, Mediów i Komunikacji Społecznej oraz Instytut Spraw Publicznych) w porozumieniu z Gminą Miejską Kraków na mocy Porozumienia o współpracy przy realizacji projektu Obserwatorium Dialogu Obywatelskiego (ODO) nr W/V/22/SO/16/2015 z dnia 5 lutego 2015 r., zawartego pomiędzy Gminą Miejską Kraków – Urzędem Miasta Krakowa a Uniwersytetem Jagiellońskim.

na weryfikację wynikającą z introspekcji czy intuicji naukowców oraz na stawiane przez nich hipotezy – oba te podejścia wzajemnie się przenikają i uzupełniają (Lewandowska-Tomaszczyk 2005). Metody komputerowego językoznawstwa korpusowego, w których używa się narzędzi NLP, są coraz powszechniej stosowane w obrębie różnych dziedzin i dyscyplin naukowych w Polsce (Świdziński 2006).

Narzędzia NLP w analizie korpusowej

Przetwarzanie języka naturalnego jest działaniem mieszczącym się na pograniczu sztucznej inteligencji i językoznawstwa. Głównym zamierzeniem NLP jest opracowanie sposobów, dzięki którym komputer przetworzy informacje tekstowe w języku naturalnym (ustne lub pisemne). Pierwotnie podstawę NLP stanowiły systemy regułowe (Littlestone 1988; Crowston, Liu i Allen 2010), następnie zaczęto stosować techniki wykorzystujące sztuczną inteligencję i modele neuronowe (Li i in. 2015; Torfi i in. 2020). Obecnie NLP rozwija się w kierunku technik hybrydowych, które integrują oba te podejścia (Dewis, Viana 2022; Khurana i in. 2022; Orlando i in. 2022; Lane i in. 2021).

Narzędzia NLP pomagają naukowcom między innymi w tworzeniu statystyk opisujących teksty, w ich automatycznej klasyfikacji oraz wyszukiwaniu informacji. Za przełom – przede wszystkim w badaniach literaturoznawczych – uważane jest przejście z *close* do *distant reading* – metodologii polegającej na zastosowaniu metod obliczeniowych do analizy danych pochodzących z dużych zasobów tekstowych (korpusów) w celu odkrycia wzorców i ukrytych w nich reguł. Metodologia ta umożliwia m.in. analizę literatury bez zaglądania do treści poszczególnych dzieł (Moretti 2005, Moretti 2013). Narzędzia NLP mają różne zastosowanie. Są przydatne w obszarze administrowania instytucjami państwowymi i prywatnymi podmiotami gospodarczymi. Wykorzystuje się je do obsługi prawnej (przeszukiwania dokumentów, automatycznego segregowania aktów itp.), a także w celu wzmocnienia ochrony cyberprzestrzeni (np. identyfikacja zagrożeń i nieprawidłowości w komunikacji elektronicznej itd.).

Modelowa analiza językowa, niezbędna przy komputerowym przetwarzaniu języka pisanego, odbywa się w kilku podstawowych etapach[5]:

[5] W języku mówionym pierwszym etapem jest analiza fonologiczna, czyli wyodrębnienie dźwięków i transkrypcja w postaci znaków tekstowych (liter). Należy zaznaczyć, że nie wszystkie z wyróżnionych etapów są możliwe do przeprowadzenia dla każdego języka, gdyż technologia dla różnych języków ma różne poziomy zaawansowania.

– rozbiór morfologiczny – dekompozycja zdań na wyrazy, a następnie wyrazów na ich części składowe (rdzeń i formanty słowotwórcze);
– analiza składniowa – identyfikacja części zdania i części mowy przy użyciu reguł gramatycznych określonych dla danego języka;
– analiza semantyczna – m.in. rozróżnienie nazw własnych od rzeczowników pospolitych, rozróżnienie typów czasowników, ujednoznacznienie znaczeń wyrazów;
– analiza pragmatyczna – uwzględnienie sensu wypowiedzi, kolokacji wyrazowych, wydźwięku emocjonalnego oraz zależności pomiędzy elementami zdania (Świdziński 2006).

Przykłady zastosowania analizy korpusowej w badaniach dyskursu

Badania medioznawcze skupiające uwagę na obserwacji i porównywaniu dyskursów tworzących sferę publiczną wymagają dzisiaj coraz częściej zastosowania narzędzi pozwalających na objęcie analizą obszernych zasobów tekstów, które stanowią możliwie najbardziej reprezentatywną próbę badawczą. Przykładami takich badań są dwa projekty realizowane w Instytucie Dziennikarstwa, Mediów i Komunikacji Społecznej, w których – we wstępnej fazie – wykorzystano elektroniczne narzędzia do automatycznego przetwarzania języka naturalnego dostępne w ramach polskiej infrastruktury naukowo-badawczej CLARIN-PL (ang. *Common Language Resources and Technology Infrastructure*) (CLARIN 2022)[6]. Pierwszy projekt objął analizę subkorpusu dyskursu parlamentarnego z lat 1989–2019 pod kątem uwarunkowań instytucjonalizacji dialogu obywatelskiego w Polsce, drugi zaś stanowił analizę porównawczą zakorzeniania się problematyki dialogu wielokulturowego w dyskursie rady miasta i dyskursie mediów w Krakowie (okresem analizy była VII kadencja Rady Miasta Krakowa, 2014–2018). Współpraca z CLARIN-PL polegała m.in. na dostosowaniu istniejących narzędzi internetowych, opracowaniu algorytmów ułatwiających realizację zaplanowanych projektów oraz przygotowaniu (ujednoliceniu) materiału do analizy. Badania zostały przeprowadzone z wykorzystaniem statystycznej analizy korpusowej i leksykalnej.

W przypadku języka polskiego technologie są na tyle rozwinięte, że wszystkie wymienione etapy analizy językowej są możliwe do przeprowadzenia.

[6] CLARIN-PL jest częścią Europejskiej Mapy Drogowej Infrastruktury Naukowej (ESFRI – European Roadmap for Research Infrastructures, European Strategy Forum on Research Infrastructures).

W obu przypadkach pierwszy etap procesu badawczego przebiegał w kilku krokach. Najpierw za pomocą aplikacji internetowej posiadającej funkcjonalności umożliwiające przeszukiwanie rozległych baz tekstowych usystematyzowano materiał badawczy według wyróżnionych dla każdego projektu słów kluczowych oraz według zdefiniowanych metadanych. Następnie metodą modelowania tematycznego (ang. *topic modeling*) rozpoznano korelację słów, określono i zobrazowano strukturę tematów (w tym stopień natężenia ich ekspozycji) poruszanych w ramach analizowanych zasobów tekstów (TOPIC 2021). W tym celu wykorzystano algorytmy *topic modelingu*, które wykonują tak zwane wnioskowanie probabilistyczne, aby określić prawdopodobną ukrytą strukturę tematyczną. Algorytmy te bazują na niezależności warunkowej w sieciach – w efekcie czego wyliczany jest wykładnik wiarygodności konkluzji skojarzonej z prawdopodobieństwem realnego zaistnienia opisanej sytuacji (Blei 2012). Jednocześnie starano się zidentyfikować tzw. terminologię dziedzinową analizowanych korpusów, posługując się narzędziem TermoPL (2021), które automatycznie wyszukuje w analizowanych zbiorach tekstów frazy językowe (rzeczownikowe) będące kandydatami na terminy badanej dziedziny (Marciniak, Mykowiecka, Rychlik 2019). Na koniec próbowano sklasyfikować wydźwięk emocjonalny badanych tekstów, wykorzystując analizę sentymentu (z ang. *sentyment analysis*) za pomocą narzędzi MultiEmo, Wydźwięk i Sentemo. Umożliwiają one identyfikację i klasyfikację fragmentów lub całych wypowiedzi ze względu na znaczenie pojawiających się w nich słów i wyrażenia nacechowane emocjonalnie (Tomanek 2014).

Organizacje pozarządowe i dialog obywatelski w dyskursie parlamentarnym

Zastosowanie narzędzi NLP w badaniach dyskursu parlamentarnego pozwoliło na scharakteryzowanie cech i zobrazowanie zmienności dyskursu decydentów politycznych dotyczącego zasad funkcjonowania organizacji pozarządowych oraz instytucjonalizacji dialogu obywatelskiego w Polsce w nowej rzeczywistości politycznej po 1989 r.

Punktem wyjścia była analiza wszystkich dokumentów (posiedzeń plenarnych obu izb, komisji Sejmu i Senatu, interpelacji poselskich, pytań i odpowiedzi) dostępnych w Korpusie Dyskursu Parlamentarnego (KDP) z lat 1989–2019 pod kątem częstotliwości występowania w nim trzech słów kluczowych: „organizacja pozarządowa", „organizacja pożytku", „dialog obywatelski". Zestawienie uzyskanych danych w podzia-

le na kadencje, lata i miesiące umożliwiło ustalenie zmienności i identyfikację okresów po 1989 r., w których następowała intensyfikacja używania przez uczestników dyskursu parlamentarnego podstawowych terminów określających trzeci sektor.

Wykres 1. Liczba wystąpień pojęć w latach 1989–2019

Graph 1. Number of keyword occurrences from 1989 to 2019

Źródło: zestawienie własne.

Na podstawie modelowania tematycznego zmapowano podstawowe wątki tematyczne i konteksty, w których analizowane terminy wystąpiły w dyskursie parlamentarnym. Badaniu poddano cztery podkorpusy wyodrębnione z zasobu KDP z okresu 1989–2019, wśród których znalazł się zbiór prac parlamentarnych zawierających wszystkie poszukiwane wyrażenia równocześnie. Z tego materiału maszynowo wygenerowano topiki (zbiory słów spójnych semantycznie), na podstawie których zidentyfikowano najważniejsze wątki tematyczne dotyczące bezpośrednio bądź pośrednio organizacji pozarządowych i dialogu obywatelskiego. Topiki te połączono w podstawowe obszary tematyczne odnoszące się do:

1) kwestii związanych z funkcjonowaniem organizacji pozarządowych w systemie politycznym (np. finansowania, samorządu terytorialnego)
2) spraw związanych z obszarami działalności i aktywności organizacji pozarządowych (np. zdrowia, nauki i edukacji czy ochrony środowiska)
3) długoterminowych polityk państwa (np. polityki mieszkaniowej, infrastruktury drogowej)
4) problemów „życia codziennego” i trosk „zwykłego człowieka”.

Rys. 1. Przykład topiku ze słowem dominującym „zdrowie”

Figure 1. Example of a topic with the dominant word "health"

Źródło: CLARIN-PL, modelowanie tematyczne.

Z tego samego materiału za pomocą narzędzia TermoPL wydobyto specyficzną dla analizowanych podkorpusów terminologię (wyrażenia dwu- i wielowyrazowe), co pozwoliło na uszczegółowienie poprzednich zestawień. W zbiorze składającym się z dokumentów parlamentarnych, w których przynajmniej raz wystąpiły wszystkie poszukiwane terminy „organizacja pozarządowa”, „organizacja pożytku” i „dialog obywatelski”, wyróżniono tzw. terminy dziedzinowe[7]. Pozyskane wyniki pozwoliły na wstępne wnioskowanie dotyczące kontekstów, w jakich posłowie i senatorowie odnosili się do spraw organizacji pozarządowych i współpracy międzysektorowej w swoich wystąpieniach.

[7] Wyróżniono i sklasyfikowano je według następujących kategorii: frekwencji występowania danego terminu w całym korpusie (F), „długości", czyli liczby wyrazów składających się na dany termin (E), frekwencji występowania danego terminu w kontekście innego terminu, czyli częstotliwości wystąpień tego terminu w zdaniach, w których pojawia się inny termin (G), liczby kontekstów (H), C-value (miary terminologicznej) częstotliwości występowania dla ekstrakcji wyrażeń, która jest wrażliwa na szczególny rodzaj terminów wielowyrazowych, tj. terminów zagnieżdżonych (D).

Tabela 1. Przykład wyrażeń wielowyrazowych odnoszących się do obszarów tematycznych „Unia Europejska" oraz „podatki i finanse"

Table 1. Example of multiword expressions related to the subjects "European Union" and "taxation and finance"

Lp.	Obszar tematyczny	Wyrażenia wielowyrazowe
1.	Unia Europejska	Unia Europejska, Komisja Europejska, państwo członkowskie, kraj Unii Europejskiej, Parlament Europejski, prawo Unii Europejskiej, Komitet Integracji Europejskiej, środek unijny, państwo członkowskie Unii Europejskiej, sprawa Unii Europejskiej, wspólnota europejska, integracja europejska, prawo unijne, środek europejski, członek Unii Europejskiej
2.	Podatki i finanse	budżet państwa, skarb państwa, finanse publiczne, podatek dochodowy, środek finansowy, rezerwa celowa, projekt budżetu, Narodowy Bank Polski, część budżetu, plan finansowy, urząd skarbowy, wydatek majątkowy, podatek akcyzowy, spółka skarbu państwa, sektor finansów publicznych, wykonać budżet, wykorzystanie środków, dopłata bezpośrednia, fundusz strukturalny, skutek finansowy, ordynacja podatkowa, dług publiczny, źródło finansowania, system podatkowy, papier wartościowy, kontrola skarbowa, wzrost cen, polityka pieniężna, wykonanie budżetu państwa

Źródło: zestawienie własne na podstawie wyników uzyskanych narzędziem TermoPL.

Sformalizowany język dyskursu parlamentarnego okazał się materiałem badawczym niezwykle trudnym do przeprowadzenia i interpretacji analizy sentymentu bez zaglądania do tekstów. Typowe słownictwo związane m.in. z procedowaniem spraw i prowadzeniem posiedzeń, określające poszczególne etapy procesu legislacyjnego itd. – mimo zastosowania tzw. stoplisty[8] – zaburzało wyniki badań. Dlatego zrezygnowano z maszynowej analizy sentymentu w projekcie, włączając ją do etapu badań *stricte* jakościowych.

Dialog wielokulturowy w dyskursie rady miasta i dyskursie mediów w Krakowie

W badaniach dotyczących dialogu wielokulturowego w Krakowie zastosowanie narzędzi NLP – według opisanego na poprzednim przykładzie schematu – umożliwiło porównanie dwóch obszernych korpusów

[8] Zestaw słów, które wyłączono z analizowanego korpusu, wśród których znalazły się m.in. formy grzecznościowe, nazwy instytucji i funkcji, słownictwo techniczne określające procedury.

tekstów, tworzących całkowicie odmienne dyskursy (instytucjonalny dyskurs radnych i dyskurs medialny). W skład pierwszego zbioru weszły stenogramy ze wszystkich sesji Rady Miasta Krakowa, które odbyły się w czasie VII kadencji Rady Miasta Krakowa (2014–2018), czyli w okresie, w którym uchwalono i zaczęto wdrażać program „Otwarty Kraków”[9]. Drugi korpus utworzono z materiałów medialnych, które ukazywały się w okresie trwania VII kadencji Rady Miasta Krakowa (2014–2018) w pięciu kanałach medialnych, każdego dnia[10].

Celem badań była próba uchwycenia tworzenia się klimatu i uwarunkowań dla budowania relacji wielokulturowych w Krakowie w sytuacji sformalizowanych działań podejmowanych przez władze miasta w tym zakresie oraz zdiagnozowania reprezentacji medialnych wielokulturowości i dialogu wielokulturowego w tym samym czasie. Starano się także ocenić poziom spójności sformalizowanego dyskursu radnych, którzy tworzą instytucjonalne ramy budowania i funkcjonowania dialogu wielokulturowego w mieście, oraz poziom dyskursu mediów, który odzwierciedla społeczne wyobrażenia i doświadczenia odnoszące się do tych relacji (Hess, Grzechnik, Zdunek 2022).

W ramach tego projektu powstał wzorowany na „korpusomacie” KDP prototyp narzędzia badawczego służącego do przeszukiwania dokumentów i analizy dyskursu Rady Miasta Krakowa VII kadencji według określonych kategorii. Podobnie jak w badaniach dyskursu parlamentarnego, wyszukiwarka ta sprawdziła się przy analizie częstotliwości, natężenia i zmienności występowania słów oraz zdefiniowanych terminów kluczowych dla podjętej w projekcie tematyki[11]. Narzędzie posłużyło w badaniach dyskursu radnych jako instrument pomocniczy zarówno w analizie ilościowej, jak i jakościowej. Dla Kor-

[9] Pierwszy długoterminowy plan dotyczący realizacji polityki otwartości na rzecz przedstawicieli mniejszości narodowych i etnicznych oraz cudzoziemców przyjęty Uchwałą nr LII/964/16 Rady Miasta Krakowa z 14 września 2016 r.

[10] Analizowano trzy portale internetowe różnego typu: serwis informacyjny, portal internetowy dziennika oraz oficjalną miejską platformę internetową. W analizie uwzględniono portal „Gazety Wyborczej” (9,1%), portal „Onet Kraków” (7,7%) oraz oficjalny serwis miejski Magiczny Kraków. Badano również „Gazetę Wyborczą Kraków” oraz „Dziennik Polski”.

[11] Punktem wyjścia dla opracowania listy słów kluczowych uczyniono dokument Program „Otwarty Kraków”, który wytyczył kierunek polityki władz miasta Krakowa w obszarze spraw związanych ze zmieniającą się – w stronę wielokulturowości – strukturą mieszkańców. Osadzenie operacyjnej definicji dialogu wielokulturowego w zakresie terminów stosowanych w dokumencie reprezentującym prawo lokalne miało na celu umożliwienie przeprowadzenia analizy porównawczej języka dokumentów z językiem mediów dotyczącym badanego zjawiska (Hess, Grzechnik, Zdunek 2022).

pusu Dyskursu Rady Miasta Krakowa (KDRMK) VII kadencji wyniki badań z wykorzystaniem modelowania tematycznego oraz identyfikacji terminów dziedzinowych okazały się niesatysfakcjonujące. Zbiór wyselekcjonowanych dokumentów był zbyt ubogi. Dodatkowym utrudnieniem był „sztuczny" język wypowiedzi i dokumentów, który jest cechą dyskursów instytucjonalnych. Wygenerowane maszynowo zbiory elementów leksykalnych były w przeważającej większości abstrakcyjne – nie były spójne semantycznie. Podobnie jak wyselekcjonowane wyrażenia wielowyrazowe, wśród których – pomimo standardowego użycia tzw. stoplisty – znalazło się wiele terminów związanych z nomenklaturą, stanowiącą nieodłączny element języka sformalizowanych posiedzeń (w tym przypadku radnych Miasta Krakowa). Tak jak w przypadku KDP, również w badaniach KRDMK zrezygnowano z maszynowej analizy sentymentu. Na etapie badań pilotażowych okazało się bowiem, że jej wyniki zniekształcają obraz nacechowania emocjonalnego dyskursu radnych. Wskazywały one, że jest on w przeważającej większości (ponad 80%) neutralny, co nie zgadzało się z wynikami analizy jakościowej.

Metody z wykorzystaniem NLP zastosowano natomiast w analizie obszernego materiału medialnego, który w pierwszej kolejności przeszukano maszynowo pod kątem występowania podstawowych słów kluczowych w postaci par lub grup słów: wielokulturowy/wielokulturowość, międzykulturowy/międzykulturowość, dialog wielokulturowy, dialog międzykulturowy. Modelowanie tematyczne (TOPIC) oraz narzędzia służące do identyfikacji terminów dziedzinowych wykorzystano do analizy korpusu dyskursu medialnego złożonego z jednostek, w których wystąpiło co najmniej jedno z poszukiwanych słów kluczowych. Wygenerowane dane pozwoliły na identyfikację i opracowanie podstawowego katalogu obszarów i kontekstów tematycznych, który został włączony – jako jedna z podstawowych kafeterii – do klucza kategoryzacyjnego stanowiącego narzędzie analizy porównawczej dyskursu rady miasta i dyskursu mediów (badań właściwych). Pomocne okazały się w tym przypadku także narzędzia do analizy sentymentu, dzięki którym udało się ustalić – w poszczególnych mediach – wstępne proporcje między materiałami nacechowanymi emocjonalnie oraz o charakterze neutralnym.

Podsumowanie

Korpusowe językoznawstwo komputerowe stanowi bardzo ważny, a niekiedy niezbędny element w metodologii współczesnych badań medioznawczych. Metody wykorzystujące narzędzia NLP umożliwiają

analizę materiałów tekstowych bez zaglądania do ich treści, co pozwala na zastosowanie skali materiału, która jest co najmniej o kilka rzędów wielkości większa niż przeciętne możliwości zespołów badawczych wykorzystujących standardowe narzędzia analizy treści.

Językoznawstwo komputerowe bardzo dobrze sprawdza się w pierwszej fazie badań. Zastosowanie narzędzi NLP pozwala na objęcie badaniem pełnej próby, precyzyjne dobranie materiału do badań właściwych (analiza jakościowa, *cloese reading*), a także opracowywanie katalogu obszarów tematycznych, który stanowi zazwyczaj podstawową kafeterię w kluczu kategoryzacyjnym – narzędziu do analizy zawartości mediów.

Badania te wymagają jednak kontynuacji i doprecyzowania metodami typowo jakościowymi na celowo dobranej próbie. Dotyczy to przede wszystkim analizy sentymentu.

Przytoczone przykłady pokazują, że dane wygenerowane za pomocą metod wykorzystujących modelowanie tematyczne oraz służących do identyfikacji terminów dziedzinowych (Walkowiak, 2017; Marciniak i in. 2019) pozwalają na wyodrębnienie wątków, rozpoznanie ich spójności, określenie i zobrazowanie struktury tematów poruszanych w analizowanych korpusach. Szczególnie przydatne staje się to w badaniach porównawczych, w których analizie poddawane są różne typy dyskursów, czego przykładem jest krakowski projekt dotyczący wielokulturowości. Udało się w nim ustalić, z jednej strony, w jakich obszarach tematycznych dialog wielokulturowy stawał się przedmiotem dyskusji radnych miasta Krakowa oraz czy, w jakim stopniu i kontekście problematyka ta była podejmowana w przekazach medialnych. Szczególnie istotnym wynikiem projektu był opis i porównanie języka – w tym terminologii – jakim posługiwali się radni i twórcy materiałów medialnych w narracjach dotyczących wielokulturowości w okresie wprowadzania przez miasto „polityki otwartości".

Literatura

Blei D.M., 2012, *Probabilistic Topic Models*, "Communications of the ACM", 55(4), http://dx.doi.org/10.1145/2133806.2133826 Blei DM.

Chomsky N., 1965, *Aspects of the theory of syntax*, MIT Press Cambridge, Massachusetts.

CLARIN, https://clarin-pl.eu/index.php/o-nas.

Crowston K., Liu X., Allen E., 2010, *Machine Learning and Rule-Based Automated Coding of Qualitative Data*, "Proceedings of the American Society for Information Science and Technology" 47, 1-2, 10.1002/meet.14504701328.

Dewis M., Viana T., 2022, *Phish responder: A Hybrid machine learning approach to detect phishing and spam emails*, "Applied System Innovation" 5, 73, https://doi.org/10.3390/asi5040073.

Hess A., Grzechnik J., Zdunek R., 2022, *Wielokulturowość w dyskursie rady miasta i dyskursie mediów. Przykład Krakowa*, Uniwersytet Jagielloński, TOC, Kraków – Nowy Targ.

Khurana D., Koli A., Khatter K. i in., 2022, *Natural language processing: state of the art, current trends and challenges*, "Multimed Tools Appl", https://doi.org/10.1007/s11042-022-13428-4.

Kleiber G., 1990, *La sémantique du prototype. Catégories et sens lexical*, Collection: Linguistique nouvelle, Presses universitaires de France, Paris.

Kučera H., Francis W., 1967, *Computational analysis of present-day American English*, Brown University Press Providence, Rhode Island.

Labov W., 1973, *Sociolinguistic patterns*, University of Pennsylvania Press, Philadelphia.

Lane H., Howard C., Hapke H., red., 2021, *Przetwarzanie języka naturalnego w akcji*, Wydawnictwo Naukowe PWN, Warszawa.

Lewandowska-Tomaszczyk B., 2005, *Powstanie i rozwój językoznawstwa korpusowego* [w:] *Podstawy językoznawstwa korpusowego*, red. B. Lewandowska-Tomaszczyk, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Li J., Chen X., Hovy E., & Jurafsky D., 2015, *Visualizing and understanding neural models in NLP*, https://aclanthology.org/N16-1082.pdf

Littlestone N., 1988, *Learning quickly when irrelevant attributes: A new linear-threshold algorithm*, "Journal of Machine Learning", 2.

Marciniak M., Mykowiecka A., Rychlik P., 2019, *TermoPL – a flexible tool for terminology extraction* [w:] *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016, 2278–2284*, red. N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis, European Language Resources Association (ELRA), *Portorož, Slovenia, 2016*, http://www.lrec-conf.org/proceedings/lrec 2016/pdf/296_Paper.pdf.

Moretti F., 2005, *Graphs, maps, trees: abstract models for a literary history*, Verso, London – New York.

Morreti F., 2013, *Distant Reading*, Verso, London – New York.

NKJP, *Narodowy Korpus Języka Polskiego*, http://nkjp.pl/ (dostęp: 20.06.2021).

Obserwatorium Dialogu Obywatelskiego, 2021, https://dialogobywatelski.org.

Ogrodniczuk M., 2017, *Lingwistyka komputerowa dla języka polskiego: dziś i jutro*, „Język Polski", XCVII(1).

Orlando G., Toledano-López J., Madera H., González A.S., 2022, *A hybrid method based on estimation of distribution algorithms to train convolutional neural networks for text categorization*, "Pattern Recognition Letters", 160.

Pang B., Lee L., 2008, *Opinion Mining and Sentiment Analysis*, "Foundations and Trends in Information Retrival", 2.

Piasecki M., 2008, *Cele i zadania lingwistyki informatycznej* [w:] *Metodologie językoznawstwa. Współczesne tendencje i kontrowersje*, red. P. Stalmaszczyk, Lexis, Kraków.

Pisarek W., 2002, *Polskie słowa sztandarowe i ich publiczność,* Universitas, Kraków.

Płaneta P., 2018, *Komputerowa analiza tekstu w dyskursach medialnych* [w:] *Metody badań medioznawczych i ich zastosowanie*, red. A. Szymańska, M. Lisowska-Magdziarz, A. Hess, Uniwersytet Jagielloński, Kraków.

TermoPL, http://www.lrec–conf.org/proceedings/lrec2016/pdf/296_Paper.pdf (dostęp: 10.10.2022).

Tomanek K., 2014, *Analiza sentymentu – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie*

danych jakościowych, „Przegląd Socjologii Jakościowej" 10(2), www.przeglad-socjologiijakosciowej.org (dostęp: 02.10.2022).

TOPIC, https://link.springer.com/chapter/10.1007%2F978–3–319–59415–6_44 (dostęp: 10.10.2021).

Torfi A., Shirvani R.A., Keneshloo Y., Tavaf N., & Fox E.A., 2020, *Natural language processing advancements by deep learning: A survey*, arXiv preprint arXiv:2003.01200.

Uchwała nr LII/964/16 Rady Miasta Krakowa 14 września 2016 r. w sprawie przyjęcia Programu „Otwarty Kraków".

Świdziński M., 2006, *Lingwistyka korpusowa w Polsce – źródła, stan, perspektywy*, „LingVaria. Półrocznik Wydziału Polonistyki Uniwersytetu Jagiellońskiego", R. 1, nr 1.

Walkowiak T., 2017, *Language Processing Modelling Notation – Orchestration of NLP Microservices* [w:] *Advances in Dependability Engineering of Complex Systems: Proceedings of the Twelfth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX*, ed. W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, J. Kacprzyk, Springer International Publishing.

Corpus Linguistics in Media Studies – a Historical and Practical Approach

Abstract

The aim of this paper is to present the benefits and risks of implementing corpus linguistics for discourse analysis. The authors describe the origins and development of Natural Language Processing (NLP) tools in a historical perspective and provide examples of their application in social sciences, particularly in the methodology of Social Communication and Media Sciences. Fragmentary findings of studies carried out at the Institute of Journalism, Media and Social Communication at the Jagiellonian University in collaboration with the CLARIN-PL consortium illustrate a practical approach to the topic. The article presents the application of NLP tools in the analysis of the corpus of parliamentary discourse from 1989-2019 in terms of determinants for the institutionalization of civic dialogue in Poland and also in the comparative analysis of multiculturalism in the city council discourse and media discourse in Krakow between 2014–2018 (7th term of the Krakow City Council). The authors indicate in which phase and at which stage of communication research the use of computational linguistics can support the conclusion.

Key words: discourse analysis, media analysis, corpus linguistics, natural language processing tools