

Magdalena Kądzioła  <https://orcid.org/0000-0002-7751-2843>
e-mail: magda.kadziola@gmail.com

Wykorzystanie metody TF-IDF w badaniu wypowiedzi mówców z różnych grup wiekowych

TF-IDF method in the analysis of utterances of Polish speakers: A comparative study of different age groups

Streszczenie

Artykuł przedstawia analizę leksyki występującej w wypowiedziach użytkowników języka polskiego należących do dwóch grup wiekowych: 18–25 oraz 65–80 lat. Celem badania jest identyfikacja i porównanie słów charakterystycznych dla obu pokoleń z wykorzystaniem metod lingwistyki korpusowej. Materiał badawczy stanowi korpus współczesnego języka mówionego liczący ponad 2,5 mln wyrazów, obejmujący teksty pochodzące m.in. z korpusu Spokes, napisów z serwisu YouTube, relacji biograficznych Archiwum Historii Mówionej oraz wystąpień sejmowych. Do analizy wybrano po 25 tekstów z każdej grupy wiekowej. Do wyłonienia leksemów charakterystycznych zastosowano miarę TF-IDF (Term Frequency – Inverse Document Frequency), obliczoną w środowisku R z wykorzystaniem pakietów tidytext i dplyr. Wyniki pokazują wyraźne różnice w warstwie leksykalnej wypowiedzi obu pokoleń. W grupie młodszej dominują elementy typowe dla mowy potocznej, takie jak wulgaryzmy, pauzy wypełnione oraz zwroty fatyczne. W wypowiedziach osób starszych częściej pojawiają się leksemy odnoszące się do przeszłości, doświadczeń biograficznych oraz relacji rodzinnych. Analiza wskazuje jednak, że różnice te wynikają nie tylko z wieku mówców, lecz także z odmiennych sytuacji komunikacyjnych reprezentowanych w badanym korpusie.

Słowa kluczowe: język mówiony, lingwistyka korpusowa, TF-IDF, leksyka polska, różnice pokoleniowe

Abstract

This article presents a lexical analysis of utterances produced by Polish speakers from two age groups: 18–25 and 65–80 years. The aim of the study is to identify and compare words characteristic of the two generations using corpus-based methods. The research material consists of a corpus of contemporary spoken Polish containing over 2.5 million tokens and including texts from several sources: the Spokes conversational corpus, subtitles from YouTube videos, biographical narratives from the Oral History Archive, and parliamentary speeches. A total of 25 texts were selected for each age group. Characteristic lexical items were identified using the TF-IDF (Term Frequency–Inverse Document Frequency) measure calculated in the R environment with the tidytext and dplyr packages. The results reveal clear lexical differences between the two groups. In the speech of younger speakers, the most prominent items include colloquial expressions, filled pauses, phatic markers and functional vulgarisms. In contrast, the speech of older speakers is characterized by words referring to the past, biographical experiences and family relations. The findings suggest that lexical differences between the groups are influenced not only by speakers' age but also by the communicative contexts represented in the corpus.

Keywords: spoken Polish, corpus linguistics, TF-IDF, lexical analysis, generational variation

Wstęp

Tradycyjnie w badaniach lingwistycznych wydziela się trzy grupy pokoleniowe:

- 1) pokolenie młode (najmłodsze) – do 30. roku życia,
- 2) pokolenie średnie – 30–60 lat,
- 3) pokolenie najstarsze (starsze) – po 60. roku życia (Dubisz, 2015: 70).

Wypowiedzi formułowane przez przedstawicieli poszczególnych pokoleń zazwyczaj się od siebie różnią, między innymi w warstwie leksykalnej (por. Kita, 2006). Przytoczony podział na grupy pokoleniowe jest umowny, bazuje na wieku metrykalnym, a granice poszczególnych kategorii mogą się zmieniać w zależności od przyjętej perspektywy badawczej (por. Zaśko-Zielińska, 2009: 118). W przedstawionych w tym artykule badaniach opieram się na wieku kalendarzowym (daty urodzenia mówców). Porównuję ze sobą słownictwo dwóch skrajnych pokoleń, a więc osób najmłodszych w korpusie z seniorami.

W literaturze przedmiotu można spotkać niekiedy rozbieżne stanowiska badaczy na temat różnic między językiem osób młodych a językiem najstarszej generacji. Niektórzy skupiający się na leksyce badacze podkreślają, że mowa ludzi młodych jest bogata w innowacje i zapożyczenia, z kolei w mowie seniorów występują słowa wychodzące już z powszechnego użycia (Gebreselassie, Godlewska, 2017: 26). Jednocześnie inni poloniści wskazują na następujące cechy wypowiedzi u osób młodych: ubóstwo środków językowych, czerpanie z angielszczyzny, humor i wulgarność (por. np. Ożóg, 2017), stosowanie strategii *mówić inaczej* i *mówić wyraziście* (por. Wileczek, 2020), u starszych zaś zauważają bogatszy zasób słów i większą kulturę słowa (por. np. Kataryńczuk-Mania, Gebreselassie, 2018) obok zjawisk takich jak: *Tip-of-the Tongue* (TOT), *Off Target Verbosity* (OTV), *Off-Topic Speech* (OTS)¹.

¹ Małgorzata Rutkiewicz-Hanczewska przedstawiła zarówno początki badań nad zjawiskiem TOT, jak i przywołała jego istotę: „Syndrom TOT pojawia się w sytuacji, gdy nie można wyszukać nazwy «własnej lub pospolitej», lecz ma się poczucie, że za chwilę to się uda” (2018: 244) (Majewska-Tworek, Zaśko-Zielińska, 2020: 177). Kolejną cechą wyróżniającą mowę seniorów jest zjawisko OTV. Po raz pierwszy zostało ono opisane w latach 80. na podstawie szeroko zakrojonych badań osób w wieku 60–90 lat (za: Świątek, 2007: 69). Zauważono, że mowa starszych osób charakteryzuje się przesadnie dużą liczbą słów oraz niewielką koncentracją na temacie wypowiedzi (OTS) (tamże, 178).

Mając do dyspozycji metody językoznawstwa korpusowego, sprawdzę, jakie cechy leksykalne wykazują wypowiedzi obu grup wiekowych, a następnie przedstawię i omówię uzyskane wyniki.

Materiał badawczy

Teksty korpusu użytego w przedstawionym badaniu pochodzą głównie ze źródeł publikowanych w internecie w ramach otwartego dostępu lub na licencji CC BY-SA 3.0, a niektóre treści zostały wykorzystane na mocy art. 27 ustawy z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych², ponadto część z nich zaczerpnęłam ze zbiorów prywatnych poddanych optycznemu rozpoznawaniu znaków (ang. *optical character recognition*, OCR) i zapisanych do plików tekstowych. Materiał badawczy w całości zebrałam i wyselekcjonowałam w latach 2016–2019, składają się zaś na niego teksty języka mówionego oraz teksty prymarnie pisane.

Na korpus użyty do badań omawianych w tym artykule składają się teksty wyekstrahowane z czterech źródeł: korpusu polszczyzny konwersacyjnej Spokes (skrótowo: SP), serwisu YouTube (skrótowo: YT), bazy Relacji biograficznych Archiwum Historii Mówionej (skrótowo: RB) oraz treści wystąpień sejmowych (skrótowo: SM). Będę odnosić się do poszczególnych typów tekstów zamiennie jako: kanał, grupa, typ lub podkorpus. Poniżej przedstawiam pokrótce poszczególne kanały.

SP

Baza danych z wyszukiwarką Spokes (Pęzik, 2015) powstała na Uniwersytecie Łódzkim w ramach polskiej infrastruktury CLARIN³. Zawiera ponad 200 godzin nagrań mowy spontanicznej przetranskrybowanej i posegmentowanej, co daje około 2,3 mln jednostek leksykalnych. Dostęp do bazy jest możliwy z poziomu przeglądarki internetowej pod adresem: <http://spokes.clarin-pl.eu>. Poza wyszukiwarką PELCRA (znaną m.in. z Narodowego Korpusu Języka Polskiego) do dyspozycji użytkowników zostaje oddany także cały zestaw metadanych mówców, takich jak wiek, wykształcenie, pochodzenie. Przy pomocy udostępnionego API⁴ (ang. *application programming interface*) badacze mogą wydobyć z bazy interesujące ich rekordy. Do tej pory przy użyciu bazy Spokes powstały liczne prace naukowe z zakresu językoznawstwa korpusowego i komputerowego (np. Wiercińska, 2018; Bocale, 2019; Ginzburg i in., 2019; Guz, 2019; Lewandowska-Tomaszczyk, 2019; Łysik, 2020). Ponadto jest ona wymieniana w artykułach przeglądowych na temat stanu badań i perspektyw lingwistyki komputerowej w Polsce (np. Pałka, Kwaśnicka-Janowicz, 2017; Ogrodniczuk, 2017; Duda, Lisczyk, 2018) obok takich zasobów jak

² Tekst jednolity Dz.U. z 2019 r., poz. 1231 z późn. zm.

³ Common Language Resources & Technology Infrastructure, <http://clarin-pl.eu/pl/o-projekcie/> [dostęp: 1.06.2023].

⁴ <http://clarin.pelcra.pl/apidocs/spokes> [dostęp: 1.03.2017].

Narodowy Korpus Języka Polskiego (NKJP) czy Słowność, co świadczy o jej wysokiej randze wśród polskich zasobów językowych.

YT

Wielu wlogerów⁵ dodaje do swoich materiałów napisy, między innymi w celu umożliwienia odbioru swoich filmów osobom niesłyszącym. Takie napisy stanowią wierne transkrypcje wypowiedzi utrwalonych w nagraniu, choć należy zaznaczyć, że w niektórych przypadkach ich autorzy pomijają na przykład omyłki jutuberów czy pauzy wypełnione. Można jednak stwierdzić, że treści takich napisów są wystarczająco szczegółowe, aby je uznać za reprezentatywne teksty współczesnego polskiego języka mówionego⁶.

Do stworzenia podkorpusu YouTube wykorzystałam treści napisów dodawanych do filmów z 34 kanałów YouTube (17 wlogerek i 17 wlogerów). Dla każdego z mówców zebrałam około jednej godziny nagrań. Niemal 74% mówców, których wypowiedzi zostały wykorzystane w podkorpusie YouTube (YT), to przedstawiciele tak zwanego pokolenia Y⁷ i pokolenia Z⁸, dla których nagrywanie wlogów stanowi naturalny sposób wyrażania siebie i komunikowania się ze swoimi widzami.

Vlogsfera to przestrzeń nie tylko rozrywkowa, lecz także komunikacyjna. Jej konstytytywnym elementem jest interaktywność, czyli wzajemne oddziaływanie na siebie podmiotów komunikacji, przykładowo w postaci reakcji na określony materiał audio-wizualny (Gorlewska, 2021: 47).

⁵ W odniesieniu do materiału pochodzącego z serwisu YouTube stosuję następujące określenia wlog, wloger, wlogerka oraz jutuber, jutuberka w pisowni spolszczonej. Obserwatorium Językowe Uniwersytetu Warszawskiego (<http://nowewyrazy.uw.edu.pl/>) notuje następujące wyrazy związane z omawianym zjawiskiem (w kolejności chronologicznej rejestracji hasła): VLOGGER (WLOGGER) – autor bloga internetowego, którego główną treścią są krótkie filmy, zwykle dotyczące aktualnych wydarzeń lub problemów związanych z różnymi dziedzinami życia codziennego, nauki, kultury itd. (rejestracja hasła: 25.07.2014); WIDEOBLOG – blog internetowy, którego zasadniczą treść stanowią pliki filmowe (rejestracja hasła: 26.09.2014); YOUTUBER (JUTUBER, rzadziej JUTIUBER) – osoba zajmująca się tworzeniem i umieszczaniem filmów w serwisie internetowym YouTube (rejestracja hasła: 5.11.2014); YOUTUBERKA (JUTUBERKA, rzadziej JUTIUBERKA) – dziewczyna lub kobieta zajmująca się tworzeniem i umieszczaniem filmów w serwisie internetowym YouTube (rejestracja hasła: 26.07.2015); WLOGOWAĆ (rzadziej WLOGOWAĆ) – tworzyć blog internetowy lub prowadzić kanał w serwisie wideo, którego główną treścią są krótkie filmy dotyczące aktualnych wydarzeń czy problemów z życia twórcy lub różnych dziedzin życia codziennego, nauki, kultury itd. (rejestracja hasła: 23.09.2016); WLOGERKA (rzadziej WLOGERKA) – autorka bloga internetowego lub prowadząca kanał w serwisie wideo, którego główną treścią są krótkie filmy dotyczące aktualnych wydarzeń czy problemów z jej życia lub różnych dziedzin życia codziennego, nauki, kultury itd. (rejestracja hasła: 5.10.2016); YOUTUBOWAĆ – WYYOUTUBOWAĆ (JUTUBOWAĆ – WYJUTUBOWAĆ) – 1) szukać filmów w serwisie internetowym YouTube, 2) zajmować się tworzeniem i umieszczaniem filmów w serwisie internetowym YouTube (rejestracja hasła: 20.10.2017).

⁶ Choć wypowiedzi osób prowadzących wlogi bywają przygotowane wcześniej, często jednak są to nagrania spontaniczne – badania z 2010 roku wykazały, że wlogerzy, prowadząc swoje kanały, zachowują się podobnie, jak podczas naturalnej rozmowy twarzą w twarz (por. Biel i in., 2011).

⁷ Pokolenie Y jest charakteryzowane również jako hashtagnation czy pokolenie selfie. Są to osoby urodzone między 1981 a 1994 rokiem, używające internetu do zdawania relacji ze swojego codziennego życia, komunikacji ze znajomymi, zakupów i wielu innych aktywności. (por. np. Stopeczyńska, 2018).

⁸ Pokolenie Z to osoby urodzone po 1995 roku, dla których „świat wirtualny i realny to ta sama rzeczywistość. Nie potrafią funkcjonować bez Internetu i mediów elektronicznych, gdyż są dla nich czymś zwyczajnym i codziennym” (Hysa, 2016: 389–390).

Możemy więc mówić o nastawieniu mówców na kontakt z widzami/słuchaczami i nieformalnym charakterze większości tych wypowiedzi. Niektóre wlogi poświęcone są jednak tematyce specjalistycznej, dlatego treści wypowiedzi z podkorpusu YT można zaliczyć do typu nieoficjalnych tekstów towarzyskich i specjalistycznych.

RB

Cyfrowe repozytorium Archiwum Historii Mówionej Domu Spotkań z Historią i Ośrodka KARTA umożliwia dostęp do ponad 5500 nagrań wywiadów biograficznych wraz z ich szczegółowymi transkrypcjami. W ramach projektu pod nazwą „Relacje biograficzne” udostępniany jest zbiór wywiadów biograficznych obejmujący tematycznie niemal cały XX wiek. Jak zaznaczają autorzy repozytorium: relacje biograficzne to linearne, subiektywne narracje o całościowym indywidualnym doświadczeniu. Świadkowie opowiadają o swoich przeżyciach od czasów dzieciństwa, okresu dorastania przez kluczowe dla nich doświadczenie wojenne po teraźniejszość⁹.

Wywiad narracyjny, który powstaje w nurcie *oral history* (por. np. Paćławska, 2008), można śmiało porównać z wywiadem indywidualnym we współczesnej etnolingwistyce (por. Bartmiński, 2008) czy dialektologii (por. np. Grochola-Szczepanek, 2006). We wszystkich wymienionych przypadkach po nagraniu materiału dokonuje się jego szczegółowej transkrypcji, a rola badacza sprowadza się do roli słuchacza. Informator własnymi słowami opowiada historię, a jego wypowiedź może zawierać omyłki czy momenty namysłu lub zawahania. Należy przy tym pamiętać, że choć wywiad narracyjny cechuje się swobodną kompozycją, to w samym założeniu różni się od spontanicznych wypowiedzi konwersacyjnych. Odróżnia go od „zwykłej rozmowy” oficjalna sytuacja komunikacyjna. Ten typ wypowiedzi można zaliczyć do spontanicznych tekstów oficjalnych (por. Majewska-Tworek, 2014). Rozmówcą informatora jest w tym przypadku badacz, który ma za zadanie, jak najbardziej ograniczyć swój udział w konstruowaniu wywodu, a informator sam nadaje bieg swojej opowieści i decyduje o jej dynamice (por. Lewandowska, 2004).

Na potrzeby omawianego podkorpusu zgromadziłam treści wypowiedzi 40 osób – 24 kobiet i 16 mężczyzn w wieku od 71 do 104 lat.

SM

Do stworzenia podkorpusu sejmowego posłużyły treści wypowiedzi posłów i posłanek z lat 2017–2018 zebrane metodą tak zwanego *web scrapingu*, czyli ekstrakcji danych na podstawie częściowo ustrukturyzowanego źródła, jakim jest strona internetowa (Maślankowski, Brzezicki, 2020: 34). W odróżnieniu od niektórych podejść do badań języka parlamentarnego (por. np. Szczyszek, 2019) nie brałam pod uwagę treści całych stenogramów z posiedzeń sejmu, ale uwzględniałam kolekcje wypowiedzi poszczególnych

⁹ Za: <https://relacjebiograficzne.pl/projekt> [dostęp: 20.12.2018].

osób, wyekscerpowane z tekstów udostępnianych na stronie internetowej Sejmu RP¹⁰. Tym sposobem zebrałam wypowiedzi 46 podmiotów – 23 kobiet i 23 mężczyzn w wieku od 33 do 73 lat. Przy pomocy wyrażeń regularnych z ich wypowiedzi usunęłam wtrącenia innych osób i określenia sytuacyjne typu „dzwonek”, „oklaski”. Rozmiary tekstów należących do tego kanału wahają się od ponad 3 tys. do ponad 25 tys. słów, a cały podkorpus sejmowy liczy ponad 509 tys. segmentów. Tematy podejmowane w wystąpieniach sejmowych dotyczą spraw publicznych i w większości reprezentują język oficjalny. Należy jednak podkreślić, że jest to reprezentacja języka oficjalnego prymarnie pisanego, choć zawiera też elementy języka mówionego. W treści występują niekiedy sformułowania potoczne, niedokończone zdania, powtórzenia czy bezpośrednie zwroty do innych osób¹¹. Z tego względu autorzy typologii NKJP zaklasyfikowali treści wystąpień sejmowych jako teksty quasimówione (por. Górski, Łaziński, 2012: 20) i w taki sposób będę się odnosić do tekstów z kanału SM.

Całkowity rozmiar omawianego korpusu przekracza 2,5 mln słów, jest więc mały¹², ale wystarczający do przeprowadzenia rzetelnych badań metodami ilościowymi (por. Eder, 2015; 2017).

Baza mówców wybranych do badania

Do przedstawionych poniżej badań potrzebne były zapisy wypowiedzi osób w konkretnych, arbitralnie założonych przedziałach wiekowych, to jest 18–25 oraz 65–80 lat¹³. Z tego powodu wybrałam z całego korpusu po 25 tekstów pochodzących od osób mieszczących się w określonych przedziałach wiekowych. Do grupy młodszej trafiły teksty z kanałów SP i YT, ponieważ są reprezentowane przez przedstawicieli najmłodszego pokolenia. Z kolei najstarsi mówcy należą do podkorpusu RB. Aby jednak uniknąć skrzywienia tematycznego w grupie starszej, dodałam do niej teksty z quasimówionego podkorpusu SM oraz jeden tekst spełniający kryteria wieku z podkorpusu YT. Określenie wieku mówców było możliwe dzięki metadanom dostępnym w bazach źródłowych (dotyczy to kanałów SP i RB) lub informacjom podanym w źródłach internetowych (odnosi się to do kanałów YT i SM). Krótkie charakterystyki obu grup przedstawiam poniżej.

¹⁰ <http://www.sejm.gov.pl/> [dostęp: kwiecień 2019].

¹¹ Przykładowo: Ile rodzina teraz musi płacić średnio? Jak ma być po zmianach? Zarzucano państwu... że nie było żadnych konsultacji co do tego projektu. A więc pytam: Z kim były przeprowadzane konsultacje? Nerwy proszę trzymać na wodzy, drodzy państwo. To nie jest tak... że wy możecie mówić wszystko, a my nie możemy mówić nic. Możemy i będziemy mówić. Z kim proponowane zmiany były konsultowane, pani minister? Jaka jest ocena tego projektu? (sejm_mazurek_beata_pos [dostęp: kwiecień 2019]).

¹² Dla porównania: NKJP ma ponad 1,5 mld słów (por. <http://nkjp.pl/> [dostęp: 18.05.2021]), a niemiecki korpus referencyjny (DeReKo) 50,6 mld słów (por. <https://www.ids-mannheim.de/en/s/corpus-linguistics/projects/corpus-development/> [dostęp: 18.05.2021]).

¹³ Należy uściślić, że ograniczam w tym badaniu pojęcie „osób najstarszych” do mówców w wieku 65–80 lat, mimo że najstarsza osoba w korpusie RB miała 104 lata, ponieważ w 2020 roku przeciętne trwanie życia mężczyzn w Polsce wyniosło 72,6 roku, natomiast kobiet 80,7 roku (Trwanie życia w 2020 roku, raport GUS: <https://stat.gov.pl/obszary-tematyczne/ludnosc/trwanie-zycia/trwanie-zycia-w-2020-roku,2,15.html> [dostęp: 27.04.2022]).

Do grupy młodszej zaklasyfikowałam 25 tekstów od 18 osób z podkorpusu SP i 7 osób z podkorpusu YT. Znalazło się tu 16 kobiet i 9 mężczyzn, a średnia ich wieku wynosiła 23 lata. Do grupy starszej wybrałam 25 tekstów pochodzących od osób z podkorpusów RB – 10 osób, SM – 14 osób i YT – 1 osoba. Do tej grupy zakwalifikowałam 10 kobiet i 15 mężczyzn, a średnia ich wieku to 70 lat. Charakterystyka statystyczna tekstów z obu grup wiekowych została przedstawiona w tabeli 1.

Tab. 1. Charakterystyka statystyczna grupy tekstów pochodzących od osób z przedziału wiekowego 18–25 i 65–80 lat (w słowach)

Grupa	Liczba tekstów	Min.	Mediana	Średnia	Maks.	Łącznie
18-25 lat	25	2271	5592	6352	15585	158789
65-80 lat	25	4364	11836	16822	57773	420555

Źródło: opracowanie własne.

Korpus użyty do prezentowanych badań jest dostępny online w serwisie Korpusomat (Kieraś, Kobylński, Ogrodniczuk, 2018) pod linkiem: <https://korpusomat.pl/corpus/4785> dla zarejestrowanych użytkowników (nazwa korpusu: *age_age*).

Metoda TF-IDF

Miara TF-IDF (ang. *Term Frequency - Inverse Document Frequency*) jest wykorzystywana w różnych narzędziach do automatycznego przetwarzania tekstów, na przykład przy generowaniu streszczeń (por. np. Christian i in., 2016) czy tak zwanej ocenie wydźwięku (ang. *sentiment analysis*) (por. np. Das, Chakraborty, 2018). Również w polskich zasobach udostępnianych w ramach infrastruktury CLARIN-PL miara TF-IDF znalazła zastosowanie w aplikacji *WebSty* (Piasecki, Walkowiak, Eder, 2018). Miara ta skonstruowana jest z dwóch członów – miary TF (ang. *Term Frequency*), która wskazuje, jak często określone słowo występuje w danym tekście, oraz miary IDF (ang. *Inverse Document Frequency*), czyli odwrotnej częstości występowania słowa w tekstach korpusu. Można więc skrótowo powiedzieć, że TF-IDF działa poprzez zmniejszenie wagi słów powszechnie używanych i zwiększenie wagi słów, które nie są używane zbyt często w zbiorze lub korpusie dokumentów (Silge, Robinson, 2016). Z formalnego punktu widzenia wartości TF-IDF oblicza się następująco:

Suppose we have a collection of N documents. Define f_{ij} to be the frequency (number of occurrences) of term (word) i in document j . Then, define the term frequency TF_{ij} to be: $TF_{ij} = f_{ij} / \max_k f_{kj}$. That is, the term frequency of term i in document j is f_{ij} normalized by dividing it by the maximum number of occurrences of any term (perhaps excluding stop words) in the same document. Thus, the most frequent term in document j gets a TF of 1, and other terms get fractions as their term frequency for this document. The IDF for a term is defined as follows. Suppose term i appears in n_i of the N documents in the collection. Then $IDF_i = \log_2(N/n_i)$. The $TFIDF$ score for term i in docu-

ment j is then defined to be $TF_{ij} \times IDF_j$. The terms with the highest TF.IDF score are often the terms that best characterize the topic of the document (Leskovec, Rajaraman, Ullman, 2010: 8–9).

W prezentowanym tu badaniu do obliczenia wartości TF-IDF dla słów w poszczególnych tekstach z obu grup wiekowych użyłam pakietów *dplyr* (Wickham i in., 2021) i *tidytext* (Silge, Robinson, 2016) w środowisku programistycznym R (por. Eder, Rybicki, Kestemont, 2016). Ponadto do szczegółowej analizy występowania wybranych leksemów korzystałam z wyszukiwarki Korpusomatu (Kieraś, Kobyliński, Ogrodniczuk, 2018; Kieraś, Kobyliński, 2021).

Przedstawienie i omówienie wyników

Osiągnięte na podstawie badanego materiału wartości TF-IDF zostały przedstawione na rysunku 1. Na wykresach zaprezentowano po 50 ważnych wyrazów dla poszczególnych grup wiekowych. Należy zaznaczyć, że prezentowane leksemy, ważne dla obu grup wiekowych, zostały poddane tylko lekkiemu czyszczeniu. Na listach stop-words¹⁴ znalazły się słowa zestawione w tabeli 2, usunięte z początków otrzymanych list wyrazowych, już po obliczeniu wartości TF-IDF dla poszczególnych słów. Decyzja o usunięciu tych konkretnych słów była umotywowana między innymi chęcią wyeliminowania powtarzających się wyrażen o tym samym brzmieniu, a różnym zapisie (np. usunięto *yy* i *y*, ale na liście zostało *yyy*; podobnie *acha* i *aha*), liczb, imion czy też zwrotów stanowiących części stałe wystąpień sejmowych.

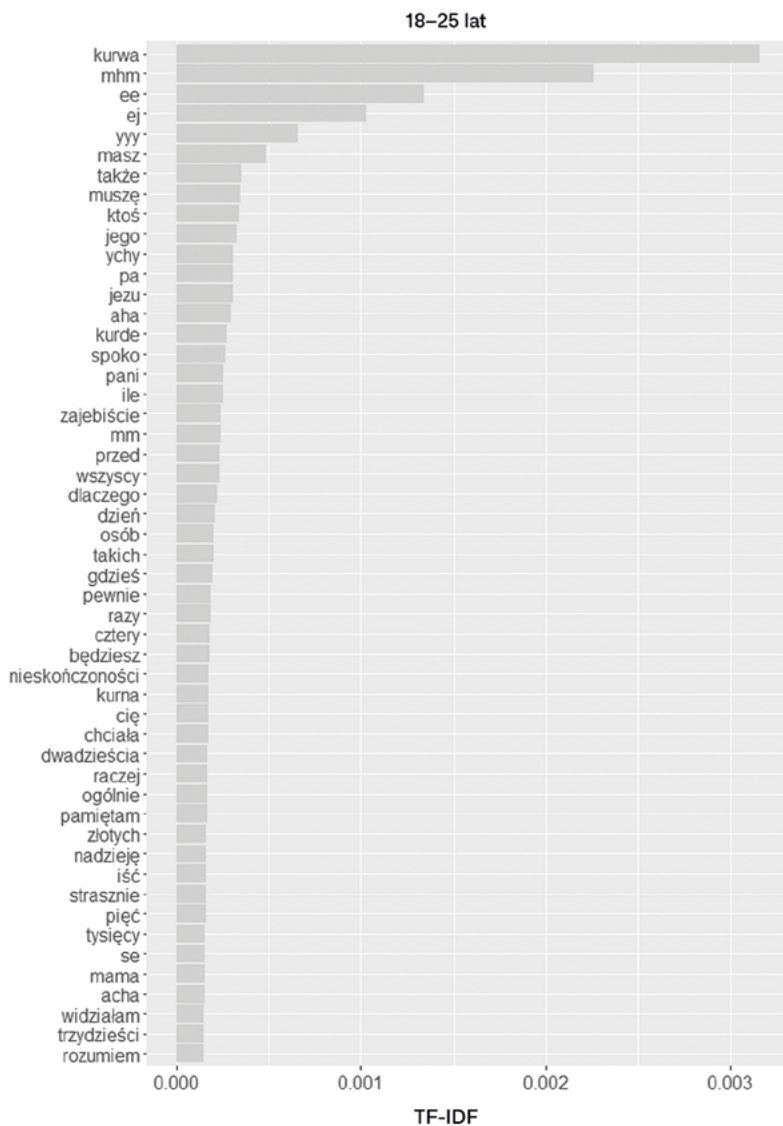
Tab. 2. Wyrazy usunięte z list wyrazów ważnych dla poszczególnych grup

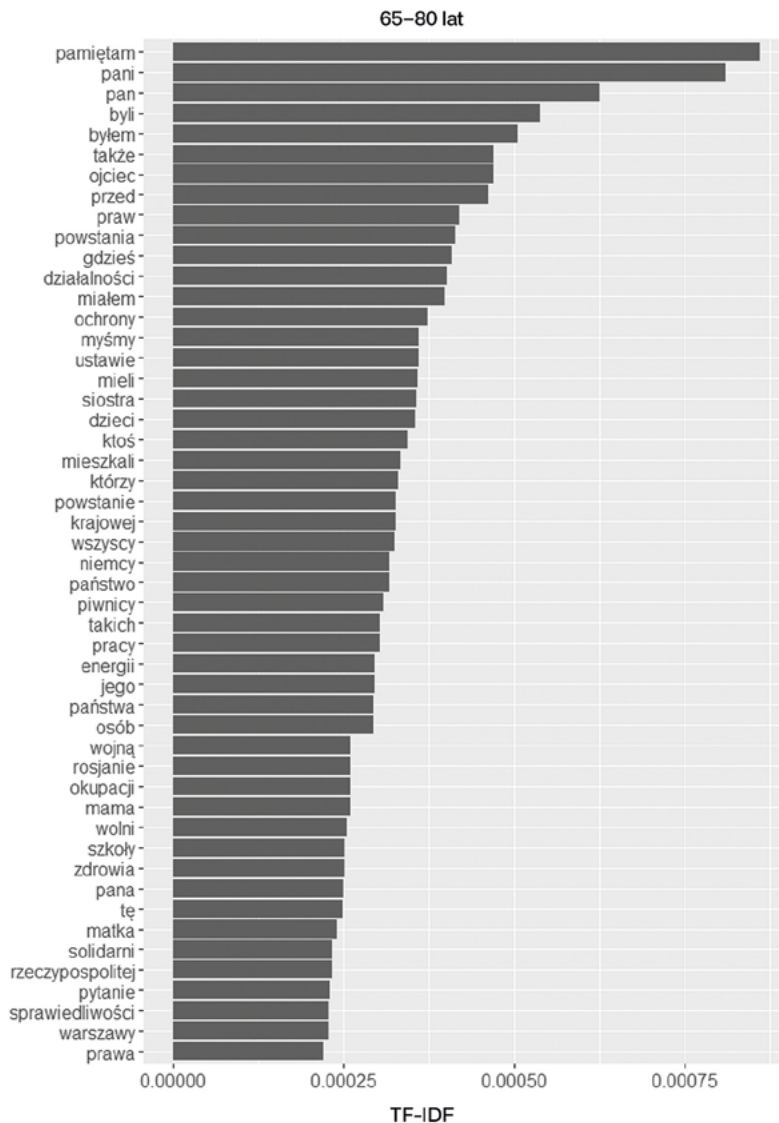
18–25 lat	65–80 lat
yy, y, iks, marcin, pani, asia, cztery, dwadzieścia, sylwia, acha, pięć, trzydzieści	wysoka, izbo, marszałku, pani, ustawy, panie, pan, ministrze, r, projekcie, sądownictwa, nr, pana, zł, ustawie, mld, 2, marszałek, 2018

Źródło: opracowanie własne.

W pierwszej kolejności zajmę się analizą poszczególnych list – grupy młodszej i grupy starszej. Następnie omówię w perspektywie porównawczej przykłady słów, które pojawiły się równocześnie na obu listach, a więc okazały się w jakimś stopniu ważne dla obu generacji.

¹⁴ Stop-words to nazwa określająca zbiór słów najczęściej występujących w danym języku (lub w danym zbiorze tekstów), tzw. słów synsemantycznych, czyli nieniosących żadnej treści informacyjnej.





Rys. 1. Wartości TF-IDF uzyskane przez poszczególne słowa ważne w obu grupach wiekowych

Źródło: opracowanie własne.

Na liście grupy młodszej zwracają uwagę przede wszystkim wulgaryzmy, pauzy wypełnione i zwroty o charakterze fatycznym, a więc słowa należące do leksykalnych wyznaczników mówioności¹⁵. Wśród wyrazów ważnych w wypowiedziach młodych

¹⁵ Wyznacznikami mówioności nazywam zestawy słów, ich bigramów i trigramów, które – według tradycyjnych opisów polszczyzny – są charakterystyczne dla spontanicznego języka mówionego. Wyodrębniam siedem ich

ludzi pojawiły się następujące funkcjonalne wulgaryzmy: *kurwa*, *jezu*, *kurde*, *zajebiście*, *kurna*, *pierdole*. Taki wynik jest zgodny z tradycyjnymi obserwacjami dotyczącymi języka młodzieży (por. Kowalikowa, 2008b; Ożóg, 2017), a także ze współcześnie notowaną ekspansją wulgaryzmów do języka ogólnego, co widać na przykładzie mediów (por. Kowalikowa, 2008a; Taras, 2011; Żmigrodzki, 2012; Sikora, 2016). Na marginesie warto odnotować, że choć na liście 50 wyrazów ważnych u grupy starszej nie znajdujemy leksemów wulgarnych, to w badanych wypowiedziach seniorów pojawiły się dwa funkcjonalne wulgaryzmy. Jeden wystąpił w wypowiedzi cytowanej przez informatora, za co mówca nawet wprost przeprosił: *Matka Boska była w jakiejś takiej grocie czy czymś wmontowana. I on mówi: – A kto to? A my mówimy, że to jest Matka Boska. – A gdzie ona? A my mówimy, że w niebie. A on mówi: – A na chuj ona tam wleza? Przepraszam, ale to fakt. To był pierwszy przejaw wojującego ateizmu, z jakim ja się zetknąłem. Z kolei drugi przypadek funkcjonalnego wulgaryzmu, jaki wystąpił w grupie starszej, to powszechne *Jezus Maria*¹⁶ stosowane ekspresywnie: *O Jezus Maria, mnie już język boli, kochani!**

Badacze współczesnej polszczyzny wskazują na brak związków między wulgarnością a wiekiem, poziomem wykształcenia czy nawet płcią¹⁷, wiążą ją zaś raczej z realizacją określonego stylu życia i używaniem tak zwanego kodu ograniczonego (por. Ożóg, 2001; Kowalikowa, 2008b, 45; Bednorz i in., 2011: 45). Co do parametru płci odmienne wnioski przedstawiają socjologowie: mężczyźni częściej od kobiet stosują wulgaryzmy w życiu codziennym (por. Mróz, Szulc, 2009). Jednak na podstawie analizowanej listy wyrazów ważnych (por. rysunek 1) można stwierdzić, że w naszym przypadku grupa młodsza była bardziej wulgarna w swoich wypowiedziach od seniorów.

Kolejnymi licznie reprezentowanymi wyrazami na liście osób w wieku 18–25 lat są pauzy wypełnione *ee*, *yyy*, *mm*¹⁸ i zwroty o charakterze fatycznym *ej*, *ychy*, *aha*. Ponadto

rodzajów: 1) ekspresywne, na przykład *uuu*, *ho ho ho*; 2) funkcjonalne wulgaryzmy, na przykład *cholera*, *boże*; 3) pauzy wypełnione, na przykład *y*, *eee*; 4) fatyczne, na przykład *m*, *aha*, *no wiecie co*; 5) bezpośrednie zwroty: a) do rozmówcy/słuchacza, na przykład *wie*, b) do rozmówców/słuchaczy, na przykład *Ślucicie*; 6) eksplicytne wskaźniki mówienia, na przykład *mówię*, *opowiem* (*ci/wam*); 7) najczęstsze słowa wspólne dla kanałów SP, RB i YT: *ja*, *no*, *ten*.

¹⁶ Por.: „Chociaż na skali między modlitwą a przekleństwem *Jezus Maria!* jest też dalekie od bieguna przekleństwa [...]. Podobnie jak okrzyk *Matko Boska!* wyraża ono nie tylko bezradność [...], ile dezorientację i niepewność [...]. W odróżnieniu jednak od *Matko Boska!* okrzyk *Jezus Maria!* sugeruje jednocześnie impuls do działania, który w połączeniu z ową niepewnością sugeruje frustrację czy irytację. Jest przy tym rzeczą charakterystyczną, że ze wszystkich wspomnianych wyrazów *Jezus Maria!* jest jedynym sformułowanym w mianowniku raczej niż w wołacz; zdenerwowany czy sfrustrowany człowiek nie zwraca się tutaj wprost do Jezusa i Marii, ale wspomina ich imiona w swoim myślowym zamieszaniu i frustracji” (Wierzbicka, 1996: 37–38).

¹⁷ Por. przykładowo: „[...] nie pojawia się – jakkolwiek zmanifestowana – korelacja między posługiwaniem się tymi słowami a wiekiem, płcią czy wykształceniem, jak również miejscem, sytuacją i scenariuszem komunikacyjnym ich użytkowników. Obserwowane przez nas sytuacje codzienne wskazują na korzystanie z tych słów niezależnie od wspomnianych wyżej czynników demograficznych czy społecznych. Jedynym obserwowalnym czynnikiem wyróżniającym jest realizacja kodu określonego stylu życia (lifestyle)” (Bednorz i in., 2011: 45).

¹⁸ Jednostka *mm* była przez Korpusomat interpretowana jako [milimetr:brev:npun], a więc skrót od wyrazu milimetr. Jednak przykłady użycia tego wyrazu potwierdzają, że występował on każdorazowo w funkcji pauzy wypełnionej czy też tak zwanego jęku namysłu, por.: *nie nie wiem o co chodzi bo ogólnie mamy jeszcze mieć przekład ustny ale to nie ale to chyba później mhm no mm mhm chodnikiem? mhm mm no uuu dziękuję dziękuję yhyhym spoko i tak mi się bardzo podoba dziękuję ci bardzo* (sp_lzpv).

wśród młodzieży pojawiły się wyrażenia potoczne lub nieoficjalne: *spoko*, *se*, *pa*, jak również bezpośrednie zwroty do rozmówcy: *masz*, *widzisz*, *daj*, *będziesz*, *cię*, *ciebie*. Wszystkie te zwroty występują w liczbie pojedynczej, a więc wskazują na bezpośredni kontakt rozmówców.

Ciekawe wyrazy, warte nieco szerszego omówienia, znajdujące się na liście grupy młodszej, to *normalnie* i *raczej*. Oba te leksemy występowały w badanych tekstach w dwóch funkcjach: 1) tradycyjnej, czyli przysłówka lub partykuły; 2) metatekstowej. Druga funkcja służy do komentowania spełnianego właśnie aktu mowy, a nie – jak standardowe przymiotniki i przysłówki – do charakteryzowania obiektów zewnętrznych w stosunku do samego procesu mówienia, to jest obiektów z rzeczywistości pozajęzykowej (por. Danielewiczowa, 2012; Maryn-Stachurska, 2019). Może też sygnalizować językową ostrożność i odpowiedzialność za słowo (Danielewiczowa, 2006: 92). Tradycyjne użycie formy *normalnie* w roli przysłówki ilustrują następujące przykłady z wypowiedzi przedstawicieli grupy młodszej: *ale już idź normalnie na koc*; *mi się wydaje że normalnie kasę dostają*; *i ona jak z mamą rozmawia to normalnie przeklina*. Z kolei na użycie drugie, częstsze w tej grupie, wskazują następujące zdania: *no coś nie wiem normalnie w sumie nawet zapomniałam*; *Marta to jest normalnie uosobienie spokoju*; *i tak to kurwa normalnie aaa to jest*. Podobnie dzieje się w przypadku formy wyrazowej *raczej*. Tradycyjnemu użyciu w funkcji partykuły, na przykład: *wiec raczej inne dni nie wchodzi w grę*; *Dobra, raczej nikt tak nie myśli*; *Smaku raczej to nie będzie miało*, towarzyszyła funkcja komentarza metatekstowego, często występującego razem z *no*: *ychy no raczej*; *no to fajnie. nasz jest... no raczej*; *o jesus nie nie nie. no raczej. kurwa nie, ja myślę że ten człowiek to powinien być hospitalizowany i to w zamkniętym ee zakładzie jakimś takim*¹⁹.

Listę grupy starszej otwierają czasowniki: *pamiętam*, *byli*, *byłem*, które można łatwo skojarzyć z podkorpusem RB, składającym się ze wspomnień i opowieści o przeszłości. Patrząc na listę wyrazów ważnych dla seniorów, da się z łatwością wskazać słowa zaczerpnięte z wywiadów biograficznych (np. *powstanie*, *okupacji*, *piwnicy*) i te z wystąpień sejmowych (np. *komisji*, *projektu*, *krajowej*), jednak są też leksemy, które mogą pochodzić z obu tych typów tekstów (np. *państwo*, *dzieci*, *szkoły*, *rzeczypospolitej*), stąd trudno je przyporządkować do któregoś z kanałów bez szczegółowej analizy wystąpień.

Wśród słów ważnych dla grupy wiekowej 65–80 lat znalazły się także nazwy członków rodziny: *ojciec*, *siostra*, *mama*, *matka*. Co ciekawe, na tej liście nie pojawia się forma *tata*. Gdy porównamy występowanie tych symetrycznych form w korpusie osób starszych, to dostrzeżemy dużą dysproporcję: wyraz *matka* występował 184 razy, *mama* – 207 razy, podczas gdy forma *ojciec* – 366 razy, a *tata* – tylko 26 razy. Nawet w obrębie jednej wypowiedzi mówcy stosowali formy niesymetryczne, na przykład: *Mało zarabiał ten ojciec, mama też*; *Ojciec przed wojną nie mógł kontynuować nauki, bo jak zmarł dziadek, to on musiał utrzymywać młodsze rodzeństwo, to znaczy, siostrę Helenę, brat chyba też, Henryk, to już pracował i jeszcze była ciocia Stasia i mama i babcia*. Z kolei forma *matka* rzadko była używana dla określenia własnej matki, a raczej posługiwano się nią w stosunku do osoby obcej, na przykład: *A ta matka jeszcze niosła tą córkę na rękę*; *Drugim*

¹⁹ We wszystkich cytatach autentycznych wypowiedzi z korpusu zachowano zapis oryginalny.

takim moim kolegą z tego podwórka był Jurek P. Ojciec jego był murarzem, matka nie pracowała, lub też sięgano po ten leksem do formułowania ogólnych stwierdzeń typu: Najgorsza matka jest lepsza od najlepszego ojca.

Interesujące wydaje się również wystąpienie formy *tę* na liście wyrazów ważnych. Intuicyjnie można założyć, że ten zaimek jest nadreprezentowany przez wypowiedzi z kanału sejmowego i dlatego znajduje się na omawianym wykazie. Jednak szczegółowa analiza występowania tego słówka w korpusie osób starszych wskazuje na powszechne użycie leksemu również przez mówców z kanału Relacji biograficznych, na przykład: *No i w tę straszną zimę, jednak powinni rodzice poczekać do wiosny; chodziliśmy tam codziennie selekcionować amunicję na tę, która była jeszcze gładka i można było załadować i strzelić, i na tę pogniecioną, a nawet jednego jutubera: są dwa sposoby na to, żeby tę sytuację zmienić; jeżeli to Cię jakoś zainspirowało, kliknij koniecznie tę łapkę, żebyś wiedział, że to Ci się podobało i że chcesz takich rzeczy więcej.* Oczywiście przywołana forma występowała też w treściach przemówień sejmowych, o czym świadczą następujące wypowiedzi: *Proszę, poprzyjcie tę poprawkę. Ona jest potrzebna tym biednym ludziom w tych biednych gminach; I wszystko wskazuje na to, że przedstawiona Wysokiej Izbie propozycja w najwyższym stopniu ma tę ważką cechę.* Warto jeszcze dodać, że w grupie młodszej wyszukiwanie [orth="tę"] dało jedynie 29 wyników, podczas gdy wśród osób starszych takich wystąpień było aż 250. Można by podejrzewać, że mamy tu do czynienia z nadreprezentacją formy *tę* spowodowaną wysoką starannością transkrypcji wykonywanych przez badaczy historii mówionej i chęcią użycia poprawnej formy słownikowej w zapisie. Jednak kontrolne zapytanie [orth="tą"] dało następujące wyniki: *była maszynka do robienia papierosów: układało się tytoń i były gilzy tak zwane, czyli to były papierosy z munsztukiem, nakładało się tą maszynkę i wciskało się ten tytoń w to; Boże kochany. Jakies takie robili, że wbijało się w rurę tą łuskę i jakoś się podpalało, i potem z wielkim hukiem ta huska wyskakiwała i leciała; jak strzelali, to krew pryskała i mnie na tą rękę napryskała, dużo krwi miałem zsiadniętej.* Na tej podstawie można zatem przyjąć, że osoby spisujące treści nagrań nie poprawiały form *tą* na *tę* w sytuacjach, gdy informatorzy rzeczywiście mówili w ten sposób.

Przykłady współwystępowania leksemów w obu grupach wiekowych

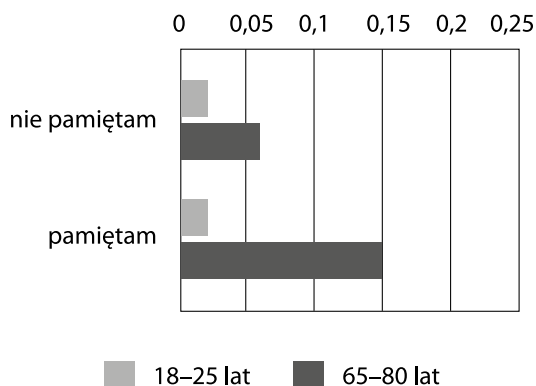
Na obu omawianych listach znalazły się jednocześnie następujące formy:

- 1) *pamiętam,*
- 2) *także,*
- 3) *mama.*

Poniżej porównuję ich użycie w poszczególnych grupach wiekowych pod względem ilościowym i ilustruję to przykładami pochodzącymi z wypowiedzi badanych mówców. Wszystkie wykresy prezentowane w tej części przedstawiają procentowy udział poszczególnych form dla korpusów grupy młodszej i starszej.

1. Pamiętam/Nie pamiętam

Przykład użycia wyrazu *pamiętam* omówię wraz z jego formą zaprzeczoną *nie pamiętam*. Porównanie występowania omawianych form w obu grupach wiekowych przedstawia rysunek 2. Widać, że zarówno *pamiętam*, jak i *nie pamiętam* były liczniej reprezentowane w korpusie seniorów. Warto zwrócić uwagę na fakt, że w grupie młodszej obie formy występowały z porównywalną częstością, z kolei wśród starszych przewagę miał wyraz *pamiętam*.



Rys. 2. Procentowe udziały form *pamiętam* i *nie pamiętam* w obu grupach wiekowych

Źródło: opracowanie własne.

Większość wystąpień formy *pamiętam* w grupie wiekowej 65–80 lat pochodzi z Relacji biograficznych (por. niżej). W przemówieniach sejmowych leksem ten pojawił się tylko raz, w wypowiedzi ministra Piotra Glińskiego: *Można różnie logikę pojmować, pamiętam to jeszcze ze szkoły*. Podobnie dzieje się w wypowiedzi wlogera Krzysztofa Litwińskiego: *Kiedy wsiadam do samochodu, włączam silnik i jadę, to zawsze pamiętam, że raz na jakiś czas muszę zrobić przegląd*. Z kolei w grupie młodszej forma *pamiętam* występowała znacząco rzadziej niż wśród osób starszych. Wybrane przykłady użycia tego wyrazu w badanym materiale przedstawiono w tabeli 3.

Tab. 3. Przykłady występowania formy *pamiętam* w obu grupach wiekowych

grupa 65–80 lat	grupa 18–25 lat
w nocy odzywały się syreny i trzeba było zbiegać do piwnicy, i znów były te modlitwy i krople walerianowe. Ja zasadniczo nie bałem się, ale zęby mi latały tak, bezwiednie, bez mojej... bezwolnie. Tak jak dziś pamiętam . Aaaaa. Nie mogłem opanować tego. (Bogdan Balcer – RB)	Szczerze mogę powiedzieć, że były to jedne z najlepszych urodzin jakie pamiętam , w sensie wiadomo, że jak byłem dzieckiem to wszystkie urodziny były super, ale odkąd pamiętam , w moim dorosłym życiu, były to naprawdę najlepsze urodziny. Dziękuję serdecznie za oglądanie. Dziękuję wam za wszystkie życzenia i oczywiście widzimy się w kolejnym filmie, w środę. Papa. (Agnieszka Grzelak – YT)

grupa 65–80 lat	grupa 18–25 lat
Bo myśmy gdzieś tam mieszkali blisko Bugu. Ja pamiętam, że do szkoły to chyba chodziłam te pierwsze dni do tak zwanych Putkowic. To jest tu zaraz za Bugiem z tamtej strony, gdzieś tam w okolicach. Bo jak się jedzie tą szosą do Drohiczyzna, pomost się mija, to tam jest drogowskaz Chorłowice i tam dalej Putkowice. To gdzieś było blisko Bugu i Niemcy kazali się odsunąć z dala od Bugu. Tak po prostu przekwaterowali ludzi, przenosili gdzie indziej. I ja pamiętam, że do tej szkoły chodziłam kilka miesięcy zaledwie. Później już pierwszej klasy nie skończyłam. Drugą zaczęłam, znowu nie skończyłam, bo takie przesiedlanie różne. (Jadwiga Czorniej – RB) Wielkanoc jak wyglądała? Wielkanocny stół to była duża rzecz. Wielkanocny stół. Te baby! Pamiętam olbrzymie baby stojące. Potem te szynki. Całe takie szynki były niepokrojone. Potem się tak kroilo po kolei. Potem, nawet pamiętam kiedyś była głowa świni, głowa prosiaka takiego, też stała. Tak, tak że pamiętam ksiądz proboszcz przychodził, pamiętam i poświęcał stół. Ale też obchodziliśmy zasadniczo w rodzinie. (Juliusz Karski – RB)	dziś kochani to wy jesteście tą gimbazą, no przecież pamiętam jak ty jeszcze chodziłeś do szkoły, pamiętam jak na lekcji opowiadałeś znajomym, że oglądałeś nowe Zapytaj Beczkę, odcinek 15, a dziś popatrz na siebie, pracujesz w korpo. (Krzysztof Gonciarz – YT)

Źródło: opracowanie własne.

Forma *nie pamiętam* była w grupie starszej mniej licznie reprezentowana niż forma niezaprzeczona. W wypowiedziach seniorów znalazł się tylko jeden przykład realizacji [orth="nie"] [] [orth="pamiętam"], mianowicie *nie bardzo pamiętam*. Pozostałe wystąpienia miały formę [orth="nie"] [orth="pamiętam"]. Ponownie w grupie wiekowej 18–25 lat występowanie tej formy było rzadsze niż wśród starszych, co może świadczyć o tym, że młodszy rozmówca nie odnosi się w swoich wypowiedziach do wspomnień i czasów minionych. Należy jeszcze odnotować, że w młodszej grupie wiekowej pojawił się tylko jeden przykład z leksemem występującym pomiędzy badanymi formami [orth="nie"] [] [orth="pamiętam"]: *nie nie pamiętam*. Wybrane zastosowania formy *nie pamiętam* w obu grupach wiekowych przedstawiono w tabeli 4.

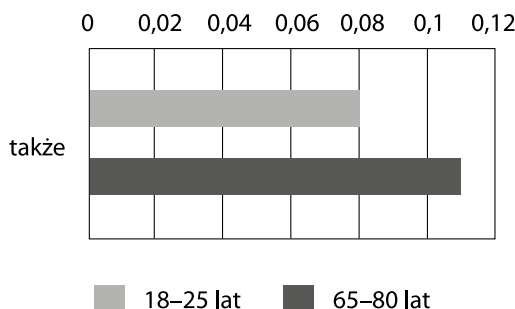
Tab. 4. Przykłady występowania formy *nie pamiętam* w obu grupach wiekowych

grupa 65–80 lat	grupa 18–25 lat
To tak 31 lat pracowałam. I potem poszłam na rentę, bo ciśnienie miałam wysokie. No i na tej rencie siedzę, i tak do tej pory siedzę. Mama żyła przy ojcu. Ojciec robotnik. Nic takiego nie było... Nie wiem, bo była wojna, a ja tak jak małe dziecko, to nie pamiętam, gdzie pracował. W jakiejś firmie, ale nie wiem gdzie. (Lucia Kowalska – RB) Za wyżywienie było tak: rano kromka chleba i nie pamiętam, chyba krążek margaryny, ale takiej podłej, i kubek czarnej kawy niesłodzonej. Chleb był jakby pieczony z takiego... cholera wie, czego, mówili, że to są trociny, ale nie czuło się trocin specjalnie. Gliniasty był bardzo, więc się na śniadanie zjadało tą skórkę, a z tego mięszu się lepilo kulki i się chowało do kieszeni, żeby w ciągu dnia żuć te kulki, oszukiwać głód. (Ryszard Szabelak – RB) To teraz warto zastosować tzw. Regułę 5 Sekund. W lutym 2017 roku Mel Robbins, prawniczka ze Stanów, opublikowała książkę „Zasada 5 Sekund”. Zasadę 5 sekund odkryła przypadkiem. Miała jakiś kryzys zawodowy straciła pracę, rozpadł się jej związek – nie pamiętam, coś bardzo poważnego. Mówiąc krótko: wpadła w depresję. Nie chciało się jej rano wstawać. Coraz trudniej było jej się zmusić do czegośkolwiek. Nie miała z życia żadnej satysfakcji, żadnej radości. (Krzysztof Litwiński – YT)	a ty nie był eś na grobach nigdzie? ja mówię nie. no to będzie tak jak mówiłem mamie jak my umrzemy to nie będzie miał nam kto świeczek stawiać. ja się tak patrzę i mówię nie nie będzie. nie nie wiem co mam kurwa powiedzieć. mam napierdalać bo akurat święto zmarłych bo normalnie nie pamiętam kurwa no normalnie nie pamiętam o zmarłych kurwa. pamięć no. no. o ja! ale tak jest. (aa5a - SP) te galerianki inne idiotki i i idioci robią ale autentycznie robią sobie imprezy typu jakieś zgadnij kto to i inne takie słoneczka kurwa ee ale mówię poważnie teraz jakieś jakieś jakieś kuźwa ee dzikie ee dzikie historie nie pamiętam. jak się nazywa ten chłopak bo to też ja ja go nie znam osobiście taki taki z lekką brodą taki dosyć wysoki który jest organizatorem zawsze tego wszystkiego (mzog - SP)

Źródło: opracowanie własne.

2. Także/Tak że

Kolejną formą wspólną dla obu grup wiekowych jest wyraz *także*. Porównanie częstości jego występowania w obu grupach wiekowych przedstawia rysunek 3. Leksem ten był liczniej reprezentowany w wypowiedziach osób z przedziału wiekowego 65–80 lat.

Rys. 3. Procentowe udziały formy *także* w obu grupach wiekowych

Źródło: opracowanie własne.

Szczegółowa analiza wystąpień prowadzi do interesujących obserwacji: podczas gdy w korpusie starszej grupy wiekowej wszystkie zastosowania formy *także* były poprawne, co znaczy, że zapisano je w transkrypcji, zgodnie ze słownikowym znaczeniem (jako *również*), to nie można tego samego powiedzieć o grupie młodszej. W zapisie wypowiedzi osób z przedziału wiekowego 18–25 lat częstszy był błędny zapis *także* w znaczeniu *tak że* niż faktyczne użycie formy *także* (w sensie *również*). Należy zatem uznać, że ta forma znalazła się na liście wyrazów ważnych dla młodzieży na skutek błędu w zapisie.

W grupie wiekowej 65–80 lat poprawne użycia *także* pojawiły się przede wszystkim w wystąpieniach sejmowych (por. rysunek 6). Ciekawy przykład, poświadczający wysoką jakość transkrypcji, znajdujemy w podkorpusie RB w wypowiedzi Bogdana Balcera.

Z kolei wśród osób w wieku 18–25 lat ponad 75% wszystkich wystąpień leksemu *także* to błędnie zapisany zwrot *tak że* (w znaczeniu *tak więc*) w funkcji spójnika. Ilustruje to rysunek 6.

Tab. 5. Przykłady występowania formy *także* w obu grupach wiekowych

grupa 65–80 lat	grupa 18–25 lat
Takie metody walki politycznej były i są przez państwa używane. Pani Marszałek! Wysoki Sejmie! Panowie, igracie z ogniem. To wy stosujecie te metody i teraz także próbujecie tę obrzydliwą, haniebną metodę wykorzystać w walce politycznej, kłamiąc na temat moich intencji i mojej wypowiedzi. (Piotr Gliński – SM)	I to jest super. Super ten film jest także ... Głupie słowo. W kontekście takiego filmu zwłaszcza, ale jest świetny pod względem samej realizacji, to znaczy jego się ogląda, jak dobrą fabułę. Bo punkty ciężkości są idealnie osadzone. Mamy punkt kulminacyjny, który prowadzi do wielkiego finału, do puenty tego wszystkiego. Więc... Nie będę mówiła konkretów bo to by były spoilery w kontekście tego, naprawdę. Ale też reżyser wiedział, co robi. (Aga Skrzypek - YT)

grupa 65–80 lat	grupa 18–25 lat
...natomiast dzisiaj na blokach, na domach pojawiają się banery terroryzujące ją, zastraszające jej rodzinę, wzywające innych do bojkotowania jej, zastraszania, a także żądające od niej, żeby opuściła Krajową Radę Sądownictwa. Mam nadzieję, że taka ustawa dla gangsterskich, nieuczciwych, chuligańskich i zastraszających celów nigdy nie będzie wydana i nikt z państwa już takiego prawa nie dostanie. Mścić się będziecie w innych okolicznościach. Na pewno nie na sędziach. (Krystyna Pawłowicz – SM) On chyba bał się przyjąć uciekiniera, dziecko uciekinierki z Warszawy. I moim losem zainteresowała się pani Kraczkiewiczowa, nauczycielka gimnazjalna, która była żoną przedwojennego profesora na Uniwersytecie Warszawskim, słynnego chemika chyba. Tak że ta także uchodziła z Warszawy, która tam niedaleko mieszkała, jakieś piekła placki na sprzedaż i była po prostu oburzona: – Jak można było dziecku odmówić? To... I powiedziała: – Ja będę go uczyć. (Bogdan Balcer)	wiesz te osoby które zawsze mają dobre oceny no to miały dobre oceny także nie wiem w ogóle nawet jak on wygląda wiesz yy mąż stary dziad młodszy jest trochę od niej no nie wiem to znaczy no kojarzę nazwisko ale w ogóle nie wiem co to za goście także się okaże yy też nie znam w ogóle nie wiem (Izabela - SP) Także moja opinia o usunięciu Lorda Kruszwiła z sieci LifeTube jest, że nie ma to znaczenia. Dziękuję bardzo, z miejsca zdarzenia, Krzysztof Gonciarz. Dobranoc państwu! Wyluzujcie, naprawdę! (Krzysztof Gonciarz)

Źródło: opracowanie własne.

Jedynie niecałe 25% wystąpień leksemu *także* w grupie młodszej to realizacje prawidłowo zapisane. Użycie formy *także* synonimicznej ze spójnikiem *również* wśród osób mających 18–25 lat poświadczają następujące przykłady:

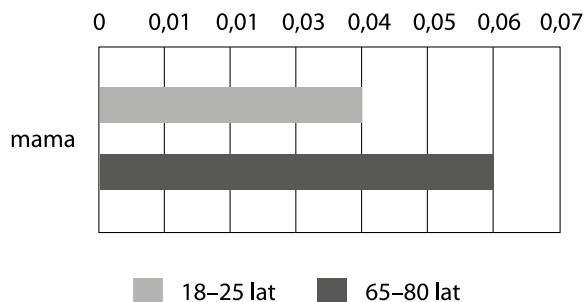
Sesja dotyczy wyspecjalizowanych dziedzin, które przynajmniej teoretycznie powinny nas interesować. A więc teoretycznie powinno nam się łatwiej do nich uczyć. Łatwiej niż na przykład wkuwać na pamięć przebieg wydarzeń z lektury, kiedy swoją przyszłość wiążemy z matematyką. Dlatego mam wrażenie, że te cechy, które różnią maturę i sesję, ale **także** sposób przygotowania do tych egzaminów będą determinowały to który z nich uznamy za trudniejszy. Bo wydaje mi się, że to nie jest sprawa obiektywna (Aleksandra Strzelecka – YT).

Obecnie wielu przedsiębiorców, **także** w Polsce, zamiast zatrudniać księgowych nabywa programy komputerowe, które w dużej mierze prowadzą za nich rachunkowość firmy (Emce Kwadrat – Huyen Pham – YT).

Na podstawie badanego materiału trudno określić, czy błędny zapis w tekstach pochodzących od najmłodszych mówców był wynikiem nieznanomości zasad ortograficznych, niechlujności językowej czy może nieumyślnej pomyłki.

3. Mama

Ostatnim przykładem współwystępowania formy na listach słów ważnych dla obu grup wiekowych jest wyraz *mama*. Na rysunku 4 przedstawione jest porównanie udziałów procentowych tego leksemu w obu badanych korpusach. Po raz kolejny omawiana forma była częstsza wśród seniorów.



Rys. 4. Procentowe udziały formy *mama* w obu grupach wiekowych

Źródło: opracowanie własne.

W grupie starszej wyraz *mama* pojawiał się tylko w treściach Relacji biograficznych. Najczęściej informatorzy określali w ten sposób własną matkę (por. wyżej). Odnosząc się do ludzi młodych, zauważymy takie realizacje, które poprzedzone były zaimkami dzierżawczymi: *twoja, moja, jego*. Przykłady występowania omawianego wyrazu w obu grupach wiekowych zaprezentowano w tabeli 5.

Tab. 5. Przykłady występowania formy *mama* w obu grupach wiekowych

grupa 65–80 lat	grupa 18–25 lat
Ale mój ojciec, taki niezamożny, zakochał się w dość takiej posażnej pannie i to było nie do przyjęcia. Wobec tego, jak w powieści, urwała się moja mama z domu, poszli gdzieś tam na stóg siana. Wróciła o określonej godzinie i była awantura, ale nie było już wyjścia. Oczywiście nie było skutków z potomstwem z tego zbliżenia, ale tak to się odbyło. Tak, tak. Oczywiście wie pani, że to były takie straszliwe, purytańskie zachowania, że na przykład – to ja już wiem z opowieści moich rodziców – moja mama miała tutaj szramę na brwi. Okazało się, że tą szramę to było tak, jak mój ojciec już przyszedł się zaręczyć – już trudno, już trzeba było to oficjalnie – to mama usiadła mu na kolanach i dziadek ją wtedy tak rąbnął, że całe życie miała tą szramę. No w każdym razie... (Alina Kałczyńska-Rodziejewicz – RB)	Karolina pojechała do biedronki i kupowała i kupiła typowo biedronkowy płyn i wszystkie klawiatury puściły z tego brudu tak na pewno no ja no no a zobacz co twoja mama przyniosła płyn do mycia podłóg nie z yyy no tak śmierdzi taką choinką nawet no no to my mamy taki tu tak je je jego mama chodzi po wszystkich yyy marketach po sklepach i wszystko kupuje co na przykład ma dzień yyy do przeterminowana się tak i ona kupuje na przykład serek danona takie jak za dwa złote są i to dzień przed terminem na przykład do końca jest trzydzieści dziewięć groszy na przykład (Apjz - SP)

grupa 65–80 lat	grupa 18–25 lat
Ja kupiłem tą bułeczkę, no i szedłem do domu. I po drodze sobie zrobiłem w niej dziurkę i tak sobie skubałem, skubałem, że, tak pamiętam, aż zwolniłem trochę, żeby jeszcze raz, jeszcze raz. No i tam nie było tego, zostało z tej bułki niewiele. Doniosłem do domu. Mama mnie ani nie zbiła, nie skrzyczała, tylko się rozplakała, jak nacisnęła i zobaczyła, że to wierzch jest, a środek nie. Ale wzięła, w tym mleku trochę co miała, rozpuściła. Po drugą nie poszedłem, nie tylko, że nie dowierzała mi, ale nie było na drugą. (Janusz Karpiński – RB) Była taka dziewczynka, może z dziewięć lat miała, z matką była. Matka była ubrana w futrze. Bo to było ciepło, bardzo ciepło, ale każdy brał to, co mógł ciepłego, bo wiedział, że nie wróci prawdopodobnie. Dziewczynka nie wytrzymała już nerwowo. Kiedy jej matkę zabili, jak oni się zbliżyli do niej po raz tam któryś, wstała i zaczęła płakać. Mówi: – Tu moja mama leży. A gdzie twoja mama? (Czesław Adamusik)	Ja doszłam do wniosku, że będę z tym żyć. I siedzę sobie przed komputerem, coś robię no i oczy mi płaczą, ja się tam przyzwyczaiłam do tego. No i wchodzi do mojego pokoju moja mama i pyta - Sylwia czemu płaczesz? - Mamo no mam alergię! - Nie, nie! To na pewno Rafał Ci coś zrobił! Co pokłóciłaś się ze swoimi znajomymi? - Mamo mam alergię! To przez alergię! Same mi oczy płaczą... - Nie! Ja Ci nie wierzę, pewnie coś ukrywasz. (Sylwia Lipka - YT) No i pewnie tak powinno być, oddzielnie, a nie. Oj, kiedyś mówiłam tak samo. A ty teraz jak? Podwójnie? Nie czy się wymienia? Podwójnie wymienia się po prostu, no to się ciesz, a nie. Ale mama daje, nie? No właśnie. No ale jak masz rachunki popłacane i wszystko... No ja wiem, tylko że widzisz, twój Michał na pewno nie je tyle, co mój Michał, bo mój Michał potrafi zjeść wszystko, co jest w domu i... No, no. (6mw5 – SP)

Źródło: opracowanie własne.

Podsumowanie

Wśród wyrazów ważnych dla grupy młodszej znalazły się formy potoczne, wulgaryzmy i zwroty wskazujące na bezpośredni kontakt z rozmówcą. Z kolei dla grupy starszej istotnymi leksemami okazały się słowa odnoszące się do przeszłości i nazwy członków rodziny. Wyniki te częściowo pokrywają się z tradycyjnymi opisami współczesnej polszczyzny mówionej (por. *Wstęp*), jednak należy stwierdzić, że na różnice w warstwie leksykalnej treści badanych wypowiedzi wpływ miało więcej czynników niż tylko wiek mówców.

Trzeba brać także pod uwagę odmienne sytuacje komunikacyjne, a więc sytuacje o różnych stopniach oficjalności, w jakich powstały teksty należące do obu podkorpusów. W grupie młodszej mamy do czynienia z nieoficjalnymi rozmowami towarzyskimi i familiarnymi (SP) oraz potocznym stylem wypowiedzi wlogerów (YT), z kolei wśród seniorów z wywiadami biograficznymi (RB) i wystąpieniami sejmowymi (SM), które zasadniczo reprezentują różne odmiany stylu oficjalnego wypowiedzi. Sytuacja komunikacyjna oddziaływała na częste występowanie wyznaczników mówioności w korpusie osób mających 18–25 lat, a w grupie starszej – form czasu przeszłego i wyrazów związanych ze sprawami publicznymi.

Omówione wyniki ukazały jeszcze jeden problem, który w badaniach korpusowych jest niebagatelny, mianowicie jakość transkrypcji. Najlepiej uwidoczniło się to podczas omawiania formy *także*, która w korpusie osób młodych w 75% wystąpień była błędnym zapisem wyrażenia *tak że*.

Na koniec wyraźnego podkreślenia wymaga jeszcze fakt, że zgromadzone na potrzeby zaprezentowanych badań teksty odzwierciedlają jedynie niewielki wycinek ogólnej współczesnej polszczyzny mówionej, a wnioski formułowane na ich podstawie trzeba przyjmować z tym zastrzeżeniem.

Bibliografia

- Bartmiński J. (2008), *O wartościach słowa mówionego*, [w:] S. Niebrzegowska-Bartmińska, S. Wasiuta (red.), *Historia mówiona w świetle etnolingwistyki*, Lublin: Wydawnictwo Polihymnia, s. 9–16.
- Bednorz M., Fąka P., Fleischer M., Grech M., Jankowska K., Siemes A., Wszolek M. (2011), *Słownik polszczyzny rzeczywistej*, Łódź: Wydawnictwo Primum Verbum.
- Biel J.I., Aran O., Gatica-Perez D. (2011), *You are known by how you vlog: Personality impressions and nonverbal behavior in youtube*, [w:] *Proceedings of the International AAAI Conference on Web and Social Media*, Barcelona, t. 5, nr 1, s. 446–449.
- Bocale P. (2019), *Deictic and epistemic distance in Polish*, [w:] I. Krapova, S. Nistratova, L. Ruvoletto, *Studi di linguistica slava Nuove prospettive e metodologie di ricerca*, Venezia: Edizioni Ca' Foscari – Digital Publishing, s. 93–110.
- Christian H., Agus M.P., Suhartono D. (2016), *Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)*, „ComTech: Computer, Mathematics and Engineering Applications”, nr 7(4), s. 285–294.
- Danielewiczowa M. (2006), *Do czego służy słowo „raczej”?*, „Polonica”, t. XXVI–XXVII, s. 83–100.
- Danielewiczowa M. (2012), *W głąb specjalizacji znaczeń. Przysłówkowe metapredykaty atestacyjne*, Warszawa: Wydawnictwo BelStudio.
- Das B., Chakraborty S. (2018), *An improved text sentiment classification model using TF-IDF and next word negation*, arXiv: preprint arXiv:1806.06407.
- Dubisz S. (2015), *Kategoria pokolenia w dialektologii*, [w:] P. Stalmaszczyk, I. Jaros (red.), *Amor verborum nos unit. Studia poświęcone pamięci Profesora Sławomira Gali*, Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 67–78.
- Duda B., Lisczyk K. (2018), *Narzędzia cyfrowe w polonistycznej dydaktyce akademickiej – zastosowania, możliwości, perspektywy*, „Forum Lingwistyczne”, nr 5, s. 143–154.
- Eder M. (2015), *Does size matter? Authorship attribution, small samples, big problem*, „Digital Scholarship in the Humanities”, nr 30(2), s. 167–182.
- Eder M. (2017), *Short samples in authorship attribution: new approach*, „Digital Humanities 2017: Conference abstracts”, Montreal: McGill University, s. 221–224.
- Eder M., Rybicki J., Kestemont M. (2016), *Stylometry with R: a package for computational text analysis*, „R Journal”, nr 8(1), s. 107–121.

- Gebreselassie J., Godlewska D. (2017), *Komunikacja językowa osób starszych*, „Język. Religia. Tożsamość”, nr 15(1), s. 21–33.
- Ginzburg J., Yusupuijiang Z., Li C., Ren K., Łupkowski P. (2019), *Characterizing the Response Space of Questions: a Corpus Study for English and Polish*, [w:] *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Stockholm: Association for Computational Linguistics, s. 320–330.
- Gorlewska E. (2021), *Wypowiedzi wartościujące jako istotny aspekt dyskursu wideoblogowego na przykładzie kanału „Stanowo” o tematyce kryminalnej*, „Białostockie Archiwum Językowe”, nr 21, s. 43–66.
- Górski R.L., Łaziński M. (2012), *Typologia tekstów w NKJP*, [w:] A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*, Warszawa: Wydawnictwo Naukowe PWN, s. 13–23.
- Grochola-Szczepanek H. (2006), *Badania fokusowe mowy mieszkańców wsi*, „Socjolingwistyka”, nr 20, s. 19–35.
- Guz W. (2019), *Direct quotation strategies in conversational Polish*, [w:] A. Bondaruk, K. Jaskuła (red.), *All around the word. Papers in honour of Bogdan Szymanek on his 65th birthday*, Lublin: Wydawnictwo KUL, s. 193–222.
- Hysa B. (2016), *Zarządzanie różnorodnością pokoleniową*, „Zeszyty Naukowe Politechniki Śląskiej”, seria „Organizacja i Zarządzanie”, nr 97, s. 385–398.
- Kataryńczuk-Mania L., Gebreselassie J., 2018, *Kompetencje komunikacyjne seniorów – wybrane wątki*, „Dyskursy Młodych Andragogów/Adult Education Discourses”, nr 19, s. 151–161.
- Kieraś W., Kobyliński Ł. (2021), *Korpusomat – stan obecny i przyszłość projektu*, „Język Polski”, CI(2), s. 49–58.
- Kieraś W., Kobyliński Ł., Ogrodniczek M. (2018), *Korpusomat – a tool for creating searchable morphosyntactically tagged corpora*, „Computational Methods in Science and Technology”, nr 24(1), s. 21–27.
- Kita M. (2006), *Komunikacja między generacjami w perspektywie stylistycznej*, „Stylistyka”, nr 15, s. 295–218.
- Kowalikowa J. (2008a), *O wulgaryzacji i dewulgaryzacji we współczesnej polszczyźnie*, „Język a Kultura”, nr 20, s. 81–88.
- Kowalikowa J. (2008b), *O wulgaryzmach w świadomości językowej młodzieży*, „Poradnik Językowy”, nr 2, s. 3–15.
- Leskovec J., Rajaraman A., Ullman J.D. (2010), *Mining of massive data sets*, Cambridge: Cambridge University Press.
- Lewandowska I. (2004), *Wywiad jako technika zdobywania informacji źródłowych w badaniu historii najnowszej*, „Echa Przeszłości”, nr 5, s. 279–299.
- Lewandowska-Tomaszczyk B. (2019), *Thumaczeniowa ekwiwalencja rozproszona: polskie coś ty i ich angielskie odpowiedniki*, „Humanitica 21”, nr 3, s. 23–38.
- Łysik M. (2020), *Wybrane funkcje pragmatyczne operatora nie wiem w języku mówionym (analiza korpusowa)*, „Prace Filologiczne”, t. LXXV(2), s. 101–114.
- Majewska-Tworek A. (2014), *Niepłynność w oficjalnej odmianie polszczyzny. Propozycja typologii*, Wrocław: Wydawnictwo QUAESTIO.

- Majewska-Tworek A., Zaśko-Zielińska M. (2020), *Mowa najstarszego pokolenia użytkowników polszczyzny. Badanie z wykorzystaniem metody językoznawstwa korpusowego – wstępne informacje o projekcie*, [w:] S. Milewski, K. Kaczorowska-Bray, B. Kamińska (red.), *Starość, język, komunikacja: nowe obszary logopedii*, Wrocław: Uniwersytet Wrocławski, s. 171–187.
- Majewska-Tworek A., Zaśko-Zielińska M., Pęzik P. (2020), *Polszczyzna mówiona miast – kontynuacja badań z lat 80. XX wieku z wykorzystaniem narzędzi lingwistyki cyfrowej*, „Forum Lingwistyczne”, nr 7, s. 71–87.
- Maryn-Stachurska D.N. (2019), *O znaczeniu przedmiotowym i metatekstowym jednostek normalny i normalnie*, „Polonica”, nr 39, s. 179–191.
- Maślankowski J., Brzezicki Ł. (2020), *Wykorzystanie mediów społecznościowych w szkolnictwie wyższym*, „Wiadomości Statystyczne. The Polish Statistician”, nr 65, s. 30–42.
- Mróz A., Szulc M. (2009), *Wulgaryzmy wśród studentów a płeć*, „Miscellanea Anthropologica et Sociologica”, nr 10(11), s. 130–143.
- Ogrodniczuk M. (2017), *Lingwistyka komputerowa dla języka polskiego: dziś i jutro*, „Język Polski”, t. I, s. 18–28.
- Ożóg K. (2001), *Polszczyzna przełomu XX i XXI wieku. Wybrane zagadnienia*, Rzeszów: Wydawnictwo Otwarty Rozdział.
- Ożóg K. (2017), *Uwagi o języku współczesnej młodzieży – między kodem ograniczonym a kodem rozwiniętym*, „Słowo. Studia Językoznawcze”, nr 8, s. 163–181.
- Paćławska E. (2008), *Zróżnicowanie gatunków mowy w tekstach historii mówionej*, [w:] S. Niebrzegowska-Bartmińska, S. Wasiuta (red.), *Historia mówiona w świetle etnolingwistyki*, Lublin: Wydawnictwo Polihymnia, s. 47–62.
- Pałka P., Kwaśnicka-Janowicz A. (2017), *Przewodnik po elektronicznych zasobach językowych dla polonistów*, „Biblioteczka TMJP”, nr 28.
- Pęzik P. (2015), *Spokes – a Search and Exploration Service for Conversational Corpus Data*, [w:] J. Odjik (red.), *Selected Papers from the CLARIN 2014 Conference*, Soesterberg: Linköping University Electronic Press, s. 99–109.
- Piasecki M., Walkowiak T., Eder M. (2018), *Open stylometric system WebSty: integrated language processing, analysis and visualisation*, „Computational Methods in Science and Technology”, nr 24(1), s. 43–58.
- Sikora K. (2016), *Kilka uwag na temat wulgaryzacji i brutalizacji polszczyzny*, „Poznańskie Spotkania Językoznawcze”, nr 32, s. 105–115.
- Silge J., Robinson D. (2016), *tidytext: Text Mining and Analysis Using Tidy Data Principles in R*, „The Journal of Open Source Software”, nr 1(3), 37.
- Stopczyńska K. (2018), *Wykorzystanie influencer marketingu w kreowaniu relacji z klientami pokolenia Y*, „Studia Oeconomica Posnaniensia”, t. 6, nr 5, s. 104–115.
- Szczyszek M. (2019), *Emocje w parlamencie parlament w emocjach: ujęcie statystyczne. O projekcie słownika polskiego parlamentaryzmu XX wieku (lata 1918–2018)*, „Prace Językoznawcze”, t. XXI, nr 3, s. 203–218.
- Świątek A. (2007), *Specyficzne zjawiska komunikacyjne w wieku senioralnym*, „Sztuka Leczenia”, t. XIV, nr 1–2, s. 69–78.
- Taras B. (2011), *Ekspansja wulgarności w języku i kulturze*, „Język Polski”, nr 5, s. 372–380.

Wickham H., François R., Henry L., Müller K. (2021), *dplyr: A Grammar of Data Manipulation*, R package version 1.0.6, <https://CRAN.R-project.org/package=dplyr> [dostęp: 20.06.2025].

Wiercińska K. (2018), *Op zoek naar perceptieverschillen bij directieftaalgebruik in het Pools en het Nederlands: een onderzoeksvoorstel*, „Acta Universitatis Carolinae Philologica”, nr 4, s. 169–178.

Wierzbicka A. (1996), *Między modlitwą a przekleństwem: O Jezu! i podobne wyrażenia na tle porównawczym*, „Etnolingwistyka”, nr 8, s. 25–39.

Wileczek A. (2020), *Kod młodzieży czy kod młodości? Społeczno-kulturowe aspekty „mediatyzacji” młodomowy*, [w:] A. Hącia, K. Kłosińska, P. Zbróg (red.), *Polszczyzna w dobie cyfryzacji*, Warszawa: Polska Akademia Nauk, s. 169–186.

Zaśko-Zielińska M. (2009), *Próba charakterystyki biolektu – język a wiek*, „Studia Linguistica”, t. XXVIII, s. 117–123.

Żmigrodzki P. (2012), *Wulgaryzmy jako problem dla leksykografa (nie tylko w związku z pracami nad Wielkim słownikiem języka polskiego PAN)*, „Annales Universitatis Paedagogicae Cracoviensis. Studia de Cultura”, nr 135, s. 123–131.