Magdalena Kądzioła[1]

[Kraków]

●●●●●●

Czynniki różnicujące wypowiedzi informatorów – analiza stylometryczna wywiadów biograficznych

Wrocławski Rocznik
Historii Mówionej
Rocznik VIII, 2018
ISSN 2084-0578
DOI: 10.26774/wrhm.206

Wstęp

Relacje między historią mówioną a językoznawstwem opisywano już m.in. z perspektywy etnolingwistyki[2], genologii[3] czy analizy tekstu i dyskursu[4]. Podobieństwo obu dyscyplin dostrzegalne jest także na poziomie metodologii – np. wywiad praktykowany w nurcie oral history i wywiad indywidulany we współczesnej dialektologii[5] cechują się podobnymi założeniami, a materiał badawczy uzyskiwany w obu przypadkach, choć służy innym celom, sprowadza się do zbliżonej formy: nagrania i zapisu treści wypowiedzi informatorów. Wraz z rozwojem językoznawstwa komputerowego pojawiają się

[1] https://orcid.org/0000-0002-7751-2843.

[2] J. Bartmiński, *O wartościach słowa mówionego*, [w:] *Historia mówiona w świetle etnolingwistyki*, red. S. Niebrzegowska-Bartmińska, S. Wasiuta, Lublin 2008, s. 9–16.

[3] E. Pacławska, *Zróżnicowanie gatunków mowy w tekstach historii mówionej*, [w:] *Historia mówiona…*, s. 47–62.

[4] D. Gocół, *Opozycja swoi/obcy w relacjach radomskiego Czerwca '76*, [w:] *Tekst – gatunek – dyskurs na przełomie XX i XXI wieku*, red. J. Szadura, Lublin 2012, s. 135–152.

[5] Por. np.: H. Grochola-Szczepanek, *Badania fokusowe mowy mieszkańców wsi*, „Socjolingwistyka", nr 20 (2006), s. 19–35.

również możliwości wykorzystania zasobów gromadzonych przez badaczy historii mówionej w nowy sposób, m.in. jako materiał do przeprowadzenia analiz metodami językoznawstwa kwantytatywnego. Cyfrowe repozytorium Archiwum Historii Mówionej Domu Spotkań z Historią i Ośrodka KARTA umożliwia dostęp do ponad 5500 nagrań wywiadów biograficznych[6] wraz ze szczegółowymi transkrypcjami. Wysoka jakość udostępnianych materiałów pozwala na ich efektywne wykorzystanie w stylometrycznych badaniach języka mówionego. Podstawowym celem analizy stylometrycznej jest określenie cech stylu autora/grupy autorów/całych nawet tradycji literackich[7], a w tym przypadku: cech stylu mówcy/grupy mówców, także w aspekcie porównawczym. Metody kwantytatywne dają wgląd w pozornie mało znaczące cechy języka, mające jednak decydujący wpływ na zróżnicowanie stylistyczne tekstów. Kilkuetapowy eksperyment stylometryczny podjęty w ramach niniejszego artykułu miał na celu sprawdzenie, które czynniki wpływają na rozróżnianie mówców. Ze względu na poszczególne cechy informatorów i tekstów, sporządzono pięć korpusów, w których różnicującymi metadanymi były: 1) płeć, 2) miejsce pochodzenia, 3) wiek, 4) długość wypowiedzi, 5) temat. W szerszej perspektywie cel ten sprowadza się do ustalenia, czy język mówiony[8] – poddany analizie kwantytatywnej, typowej dla badań nad literaturą – podlega podobnym prawom, co teksty literackie[9].

---

[6] Por. opis kolekcji: „AHM to największy w Polsce zbiór wywiadów biograficznych [...] obejmujący tematycznie niemal cały XX wiek"; „Wybrane do portalu relacje to świetne przykłady wywiadów narracyjnych i biograficznych – trwają wiele godzin, rozmówca samodzielnie prowadzi opowieść o swoim życiu, a pytania do niego pojawiają się dopiero pod koniec nagrania", https://relacjebiograficzne.pl/projekt (dostęp: 1 IX – 20 XII 2018 r.).

[7] Por.: M. Eder, *Metody ścisłe w językoznawstwie i pułapki pozornego obiektywizmu. Przykład stylometrii*, „Teksty Drugie", nr 2 (2014), s. 90–105.

[8] Język mówiony traktowany jest w niniejszym artykule jako odmiana języka ogólnego; „Język pisany i mówiony traktowane są jako dwa różne podsystemy języka etnicznego, co zostało wyraźnie stwierdzone w wydanej w roku 1968 *Praktycznej stylistyce* Anny i Piotra Wierzbickich, a następnie w wielu innych pracach językoznawczych. Określa się je również jako dwie odmiany języka ogólnego, przy czym odmianę mówioną nieoficjalną (nieformalną, spontaniczną, codzienną) nazywa się również odmianą potoczną w opozycji do odmiany oficjalnej (starannej, opracowanej, formalnej)", J. Labocha, *Pragmatyczne mechanizmy składni języka mówionego*, „Slavia Occidentalis", nr 69 (2012), s. 139.

[9] Por. np.: M. Eder, *Style-markers in authorship attribution: a cross-language study of the authorial fingerprint*, „Studies in Polish Linguistics", nr 6 (2011), s. 99–114.

Dodatkowy cel niniejszego artykułu to sprawdzenie skuteczności samej metody w zastosowaniu do nowego – z punktu widzenia tradycyjnej stylometrii – materiału badawczego: tekstów języka mówionego[10], które różnią się od tekstów literackich np. długością[11]. Podejmowano już badania metodami językoznawstwa kwantytatywnego m.in. na materiale wystąpień sejmowych czy mów prezydenckich[12], jednak wymienione teksty należą do odmiany oficjalnej języka mówionego, a nierzadko stanowią po prostu odczyt wcześniej przygotowanego tekstu pisanego. Treści relacji biograficznych można z kolei zaliczyć do odmiany nieoficjalnej języka mówionego: informator własnymi słowami opowiada historię, jego wypowiedź zawierać może omyłki, momenty zawahania czy nielinearne struktury składniowe, typowe dla mowy spontanicznej. Należy przy tym pamiętać, że choć wywiad narracyjny cechuje się swobodną kompozycją[13], to jednak w samym założeniu różni się od spontanicznych wypowiedzi codziennych rozmów. Rozmówcą informatora jest w tym przypadku badacz, który ma za zadanie jak najbardziej ograniczyć swój udział w konstruowaniu wywodu[14], a informator sam nadaje bieg swojej opowieści i decyduje o jej dynamice. Powstałe teksty, podobnie jak dzieła literackie konkretnych autorów, powinny zawierać więc zestawy cech charakterystycznych dla każdego z mówców/grup mówców, możliwe do zbadania metodami ilościowymi.

---

10 Por. np.: J. Labocha, *op. cit.*, s. 139–145; P. Pęzik, *Język mówiony w NKJP*, [w:] *Narodowy Korpus Języka Polskiego*, red. A. Przepiórkowski, M. Bańko, R. Górski, B. Lewandowska-Tomaszczyk, Warszawa 2012, s. 37–48.

11 Por. np. rozmiar korpusu: M. Eder, J. Rybicki, K. Młynarczyk [et al.], *1000 Novels Corpus*, CLARIN-PL digital repository, 2016.

12 Np. J. Herz, A. Bellaachia, *The authorship of audacity: Data mining and stylometric analysis of Barack Obama speeches*, *Proceedings of the International Conference on Data Mining (DMIN)*, (DMIN), b. m. 2014.

13 „Wywiad narracyjny nie zakłada stawiania wstępnych hipotez, by nie ukierunkowywać rozmowy. Stosuje się tu tak zwaną *zasadę otwartości* polegającą na tym, że to badani w znacznej mierze określają ostateczne kontury przedmiotu badania", I. Lewandowska, *Wywiad jako technika zdobywania informacji źródłowych w badaniu historii najnowszej*, „Echa Przeszłości", t. V (2004), s. 279–299.

14 Zob. np. I. Lewandowska, *op. cit.*

Korpus

Na korpus użyty w ramach niniejszego artykułu składają się teksty pochodzące z trzech wybranych kategorii tematycznych z repozytorium www.relacjebiograficzne.pl[15], tj.: „Warszawa", „wieś" i „ziemiaństwo". Kategorie zostały wybrane spośród dwunastu wyróżnionych w bazie[16], przede wszystkim ze względu na wspólny zakres tematyczny związany z miejscami zamieszkania i sposobem życia[17]. Na podstawie nagrań wykluczono z korpusu teksty, których autorzy nie posługiwali się płynnie językiem polskim. W rezultacie zgromadzono 41 tekstów o długościach między 7267 a 60 435 słów – taka rozpiętość wynika przede wszystkim z różnic w czasie trwania wywiadów oraz indywidualnych różnic w sposobie mówienia każdego z informatorów. Całkowity rozmiar korpusu przekracza 1,2 mln jednostek leksykalnych. Pytania, dopowiedzenia i inne wypowiedzi badaczy przeprowadzających wywiady zostały usunięte z treści wszystkich zgromadzonych relacji biograficznych.

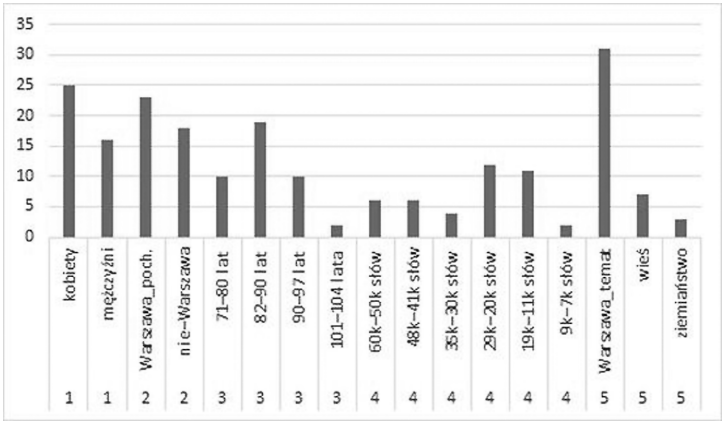
Treści 41 wywiadów biograficznych zostały wykorzystane w każdym z pięciu korpusów. Próbki tekstowe podzielono na różne klasy, w zależności od badanej cechy. I tak, w korpusie pierwszym sprawdzano wpływ płci na klasyfikację tekstów – w tym celu wydzielono dwie klasy: kobiety (A) i mężczyźni (B); w korpusie drugim – miejsce pochodzenia – dwie klasy: Warszawa (A) i nie-Warszawa (B); w korpusie trzecim – wiek – cztery klasy: 71–80 lat (A), 82–90 lat (B), 90–97 lat (C), 101–104 lata (D); w korpusie czwartym – długość wypowiedzi – sześć klas: 60k–50k słów (A), 48k––41k słów (B), 35k–30k słów (C), 29k–20k słów (D), 19k–11k słów (E), 9k–7k słów (F); z kolei w korpusie piątym – temat – trzy klasy: Warszawa (A), wieś (B), ziemiaństwo (C).

[15] Dostęp: 1–10 IX 2018 r.

[16] Pozostałe kategorie tematyczne to: „obozy koncentracyjne", „Kresy", „wojsko", „cudzoziemcy w PRL", „zesłania i łagry", „powstanie warszawskie", „konspiracja powojenna", „praca przymusowa", „Holokaust".

[17] Należy tu wyjaśnić, że w podobnej kategorii, „Kresy", znajdują się wywiady zaklasyfikowane równocześnie do innych kategorii, w tym „wieś" i „ziemiaństwo", dlatego poprzestano na trzech wymienionych wyżej kategoriach.

Rys. 1. Liczba informatorów w poszczególnych klasach dla pięciu korpusów.



W eksperymentach nie wykorzystano bezpośrednio plików tekstowych zawierających treści pochodzące od każdego z informatorów, zamiast tego teksty należące do poszczególnych klas zostały najpierw połączone w jeden duży dokument (osobny dla każdej klasy), który następnie podzielono na 15 równych, wyłonionych losowo, próbek. Dało to równą ilość próbek tekstowych dla każdej klasy w poszczególnych korpusach; naturalnie próbki różniły się długościami. Zabieg ten miał pozwolić na uniknięcie wpływu autorstwa na wyniki[18], ponieważ badano nie cechy wyróżniające poszczególnych mówców, a cechy całych grup mówców, np. kobiet w porównaniu z mężczyznami.

Przed przeprowadzeniem serii eksperymentów zbadano najczęstsze słowa (MFW – *most frequent words*) oraz ciągi słów: bigramy (dwa słowa występujące w tekście obok siebie) i trigramy (trzy słowa występujące w tekście obok siebie) w różnych pozycjach na liście frekwencyjnej. Celem takiego postępowania było – po pierwsze – wyznaczenie pozycji, w których występują najbardziej interesujące nas jednostki leksykalne oraz – po drugie – ukazanie zasadniczych różnic między wyrazami na różnych pozycjach

[18] Por. J. Mandravickaitė, T. Krilavičius, *Stylometric Analysis of Parliamentary Speeches: Gender Dimension*, [w:] *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, red. T. Erjavec [et al.], Valencia 2017, s. 102–107.

listy frekwencyjnej. Na początku listy znajdują się tzw. wyrazy funkcyjne – jest to zjawisko typowe nie tylko dla języka polskiego[19], a im dalsza pozycja na liście, tym więcej wyrazów znaczących. Poniżej (tab. 1) zilustrowano różnice pomiędzy jednostkami leksykalnymi w zależności od ich pozycji na liście, według wybranych progów.

Tabela 1. Dwadzieścia najczęstszych: słów, bigramów i trigramów dla całego korpusu.

Pozycje 1–20 na liście frekwencyjnej		
to	to było	i tak dalej
i	nie było	w każdym razie
w	no i	nie wiem czy
nie	w tym	w tej chwili
się	to był	że to jest
na	to jest	to nie było
że	tak że	z tym że
z	i tak	jak to się
do	nie wiem	w ten sposób
tak	bo to	okazało się że
tam	to była	w tym czasie
było	to już	że tak powiem
ja	to nie	już nie pamiętam
jak	że to	w pewnym momencie
bo	to się	tak że to
a	i to	w ogóle nie
ale	w ogóle	i to było
był	i tam	ja nie wiem
już	do tego	bo to było
była	w tej	to było w
Pozycje 100–120 na liście frekwencyjnej		

[19] Por. np.: J.F. Burrows, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, Oxford 1987.

ogóle	przed wojną	tam nie było
która	do warszawy	wyszła za mąż
są	i wtedy	po jakimś czasie
mam	bo nie	w stosunku do
jakiś	ja byłam	bo to już
oczywiście	w domu	nie pamiętam jak
które	tam w	to było coś
Niemcy	było tak	nie było w
ich	że jak	nie pamiętam czy
jako	tam się	się okazało że
ci	i ten	ja już nie
dużo	tam na	tam w tym
zawsze	bo tam	wiem czy to
mój	ale w	z jednej strony
nic	do domu	do domu i
kiedy	że tak	i na tym
można	były takie	i to wszystko
lat	i ona	no i ja
być	nie nie	do tej szkoły
moja	z tych	i pamiętam że
dzieci	się to	no i to
Pozycje 1000–1020 na liście frekwencyjnej		
dzieckiem	jej mąż	tak że było
gospodarstwo	mnie tam	tak że te
jeździć	mojej matki	tej chwili jest
kontakty	mówi no	to była duża
kościele	nie mamy	to było jeszcze
krzyż	on tak	to ja pamiętam
mogło	się jakoś	to już wtedy
Niemca	tak bo	to się odbywało
Niemcami	tym to	to tam była
pociągu	w ciągu	to też nie
same	ale było	to to była
stary	ale później	to to było
wyobrazić	bardzo się	tylko po prostu
wyszłam	były tak	tym że nie
zdjęcie	co jest	w domu nie
zwłaszcza	dla tych	w ogóle i
żoną	do roku	wiem co się
człowiekiem	domu to	więc to była
jeździłem	gdzie była	wszystko było w
każdej	jakiś taki	z tego to
lekcje	nie mogłem	z tym co

Analiza listy frekwencyjnej umożliwia wydobycie znaczników języka mówionego – najlepiej ilustrują to trigramy z pierwszej dwudziestki; są to uniwersalne, niezależne od tematu rozmowy, stałe elementy pojawiające się w wypowiedziach, tj.: „i tak dalej", „w każdym razie", „że tak powiem", „tak że to"[20]. Z kolei za charakterystyczne dla gatunku wywiadu biograficznego można uznać zwroty: „już nie pamiętam", „nie pamiętam jak", „nie pamiętam czy", „i pamiętam że", „to ja pamiętam", pojawiające się wśród trigramów na różnych pozycjach listy. Warto zauważyć, że pośród MFW i bigramów w porównywanych zakresach nie znalazły się żadne jednostki związane z „pamiętaniem". Leksem „pamiętam" po raz pierwszy pojawia się dopiero na 55. miejscu listy frekwencyjnej MFW, zaś wśród bigramów – „nie pamiętam" – na 35. miejscu.

Słowa pojawiające się poniżej pozycji 1000. na liście frekwencyjnej MFW odnoszą się już do samej treści relacji biograficznych: znajdziemy tu jednostki związane zarówno z dzieciństwem („dzieckiem", „lekcje"), religią („kościele", „krzyż"), jak i zapewne z okupacją („Niemiec", „Niemcami"). W tym przedziale znajdują się już same wyrazy znaczące. Warto zaznaczyć, że wszystkie eksperymenty przeprowadzane były na niezlematyzowanym materiale, tzn. wszystkie wyrazy pozostawiono w formach fleksyjnych, w których występowały oryginalnie w tekstach (stąd obecność różnych form tego samego wyrazu, np. „jeździć" – „jeździłem", „Niemiec" – „Niemcami").

Metoda

Metodologiczną podstawą przeprowadzonych w niniejszym studium eksperymentów jest założenie, że frekwencje kilkudziesięciu czy kilkuset badanych najczęstszych wyrazów układają się w profil frekwencyjny (w stylometrii określa się go czasem mianem stylomu albo stylistycznego odcisku palca), który różni się drobnymi, prawie niezauważalnymi detalami pomiędzy różnymi próbkami tekstowymi. Istnieje szereg tzw. wielowymiarowych metod, dzięki którym większa liczba owych z pozoru pomijalnych różnic mieści się w sumarycznej różnicy pomiędzy badanymi tekstami. Co więcej, różnice między badanymi tekstami na ogół nie są takie same, dzięki czemu za pomocą metod wielowymiarowych daje się wyodrębnić całe grupy tekstów do siebie podobnych i znacząco różniących się od innych grup tekstów.

[20] Por. P. Pęzik, *op. cit.*

Podstawową miarą w badaniach stylometrycznych jest delta Burrowsa[21], algorytm służący do obliczania odległości między wyrazami na listach frekwencyjnych poszczególnych elementów korpusu. Jej zmodyfikowaną wersją jest delta Edera, którą uznaje się za efektywniejszą przy badaniach nad językami fleksyjnymi[22], dlatego też w prezentowanych badaniach użyto tej drugiej. Wszystkie eksperymenty przeprowadzono za pomocą pakietu *stylo*[23] w środowisku programistycznym R. Na podstawie wstępnej analizy trzech wyżej opisanych wskaźników stylu (MFW, bigramy, trigramy) zdecydowano się na przeprowadzenie dalszych badań na MFW i trigramach, rozpoczynając od 100–1000 najczęstszych jednostek leksykalnych, następnie wykonywano próby: 200–500, 300–1000, 500–1000, kolejno: od pierwszej pozycji na liście, od setnej i od tysięcznej. Wyniki wszystkich eksperymentów przeprowadzonych na MFW były zbliżone, podczas gdy dla trigramów wyniki były niestabilne. Dlatego, z uwagi na najbardziej klarowne efekty, uwzględniając także reprezentację graficzną, niżej omówione są rezultaty otrzymane dla 500 MFW, licząc od tysięcznej pozycji na liście frekwencyjnej.

Aby zaprezentować graficznie wyniki przeprowadzonych eksperymentów, wykorzystano dwa sposoby: skalowanie wielowymiarowe (MDS – *multidimentional scaling*) i analizę skupień (CA – *cluster analysis*). Obie metody biorą pod uwagę różnice między poszczególnymi frekwencjami słów: za pomocą MDS przekształca się je w odległości pomiędzy punktami na wykresie w przestrzeni kartezjańskiej, w przypadku CA – na gałęziach drzewa. Bliskość punktów mówi nam o podobieństwie tekstów reprezentowanych przez te punkty.

Wyniki

Poniżej opisano otrzymane wyniki dla każdego z pięciu korpusów.

1. Płeć

Płeć autora jest cechą, która w literackich badaniach stylometrycznych zazwyczaj jest możliwa do rozróżnienia. Widoczne są wyraźne różnice między językiem autorów a językiem autorek powieści anglojęzycznych, zwłaszcza

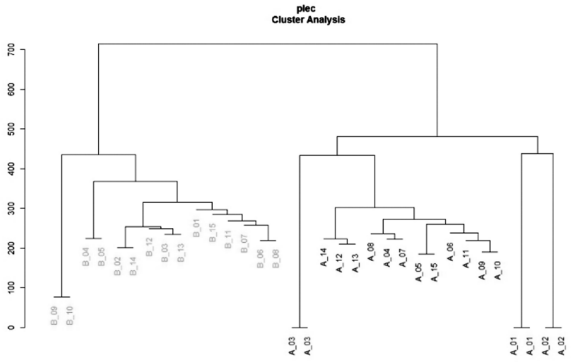
[21] Por.: J.F. Burrows, *„Delta": A Measure of Stylistic Difference and a Guide to Likely Authorship*, „Literary and Linguistic Computing", nr 17 (2002), s. 267–287.

[22] Por.: J. Mandravickaitė, T. Krilavičius, *op. cit.*

[23] M. Eder, M. Kestemont, J. Rybicki, *Stylometry with R: A package for computational text analysis*, „R Journal", nr 16 (1) (2016), s. 107–121.

w literaturze współczesnej[24]. W przypadku przemówień parlamentarnych kobiet i mężczyzn w języku litewskim efektem analizy stylometrycznej była klasyfikacja tekstów według płci[25]. W serii naszych eksperymentów było podobnie – korpus pierwszy dawał zawsze najlepsze wyniki, tzn. w każdym z wariantów przeprowadzonego eksperymentu teksty dzieliły się wyraźnie na dwie grupy, według płci mówcy. Można to zaobserwować na wykresach – na obu widać formowanie się dwóch wyraźnych podgrup: na dendrogramie analizy skupień (rys. 2) próbki podzieliły się na dwie główne gałęzie drzewa, na wykresie skalowania wielowymiarowego (rys. 3) próbki ułożyły się w postaci dwóch skupisk o wyraźnie zaznaczonych brzegach (tzn. grupy nie zachodzą na siebie). Nie należy jednak zapominać, że w naszym przypadku eksperymenty odbywały się na materiale niezlematyzowanym.

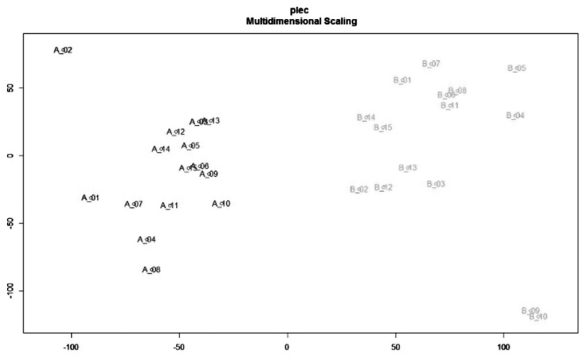
Rys. 2. Różnice między kobietami (prefiks A) a mężczyznami (prefiks B). Analiza skupień.



[24] Np.: J. Rybicki, *Vive la différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies*, „Digital Scholarship in the Humanities", nr 31 (4) (2016), s. 746–761; S.G. Weidman, J. O'Sullivan, *The limits of distinctive words: Re-evaluating literature's gender marker debate*, „Digital Scholarship in the Humanities", nr 33 (2) (2017), s. 374–390.

[25] J. Mandravickaitė, T. Krilavičius, *op. cit.*

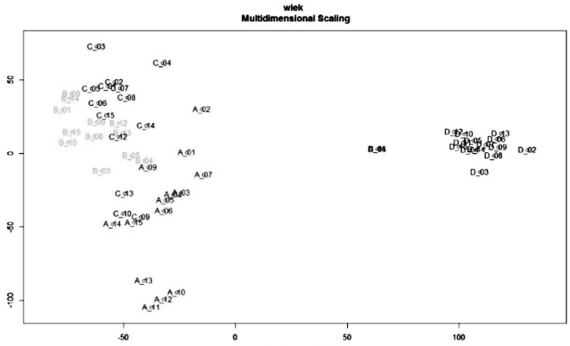
Rys. 3. Różnice między kobietami (prefiks A) a mężczyznami (prefiks B). Skalowanie wielowymiarowe.



2. Miejsce pochodzenia

Korpus drugi składał się tylko z dwóch klas: „Warszawa" i „nie-Warszawa". Taki podział spowodowany był faktem, że informatorzy urodzeni w Warszawie stanowili aż 56% badanej grupy, a miejsca urodzenia pozostałych mówców rozproszone były od Białegostoku przez Łódź, po Kielce i Lwów. W tym przypadku więc hipoteza badawcza brzmiała: mówcy urodzeni w Warszawie będą wykazywać się podobieństwami i zgrupują się w opozycji do mówców nieurodzonych w Warszawie. Graficzna prezentacja wyników pokazuje, że nie udało się dokonać jednoznacznego podziału na dwie klasy, mimo to można zaobserwować mniejsze skupienia, w których mówcy się grupowali (rys. 4), np. grupa dziewięciu mówców z klasy B, osobna gałąź trzech mówców z klasy A. Jednak miejsce urodzenia, które potraktowano jako potencjalną cechę różnicującą, nie zawsze jest tożsame z miejscem zamieszkania, w którym informator nabierać musiał różnych nawyków związanych z wypowiadaniem się. Należy zatem sądzić, że właśnie to miało wpływ na otrzymany podział. Udało się więc ustalić, że deklarowane miejsce urodzenia nie jest cechą, która różnicuje mówców w badaniach stylometrycznych prowadzonych na niewielkim korpusie.

Rys. 4. Różnice między informatorami urodzonymi w Warszawie (prefiks A) a informatorami urodzonymi w innych miejscowościach (prefiks B). Analiza skupień.



3. Wiek

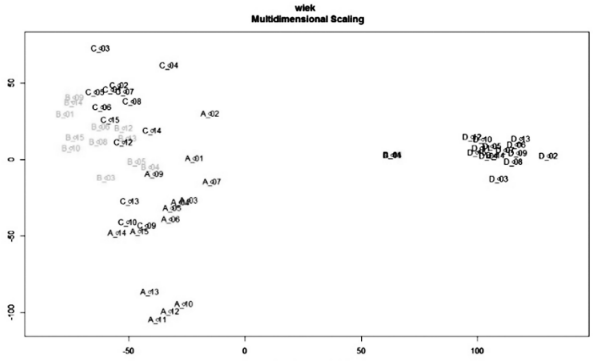
Badanie zależności stylu od wieku mówcy jest interesujące z co najmniej dwóch powodów. Po pierwsze, język – jako zjawisko społeczne – ewoluuje i te zmiany są w jakimś stopniu widoczne również w sposobie mówienia osób starszych, urodzonych wiele dziesięcioleci temu. Po drugie, istnieją badania przeprowadzone na tekstach literackich pokazujące, że język jednostkowy sam w sobie zmienia się w czasie. Niektórzy pisarze z wiekiem dopracowują swój styl, czego dowodzi przykład Henry'ego Jamesa[26], u innych zaś widać dezintegrację funkcji poznawczych i upraszczanie języka, spowodowane na przykład demencją[27]. Pytanie, którego dotychczas jeszcze nie postawiono, odnosi się do zróżnicowania języka mówionego osób w różnym wieku. W naszym przypadku wiek mówców wahał się od 71 do 104 lat, a informatorzy zostali podzieleni na klasy ze względu na swój wiek w momencie nagrania co około 10 lat. Korpus trzeci składał się z czterech klas, w tym najliczniejszą stanowili mówcy w przedziale wiekowym 82–90 lat (19 osób), zaś najmniej liczną – mówcy powyżej stu lat (tylko 2 osoby). Ta duża dysproporcja w liczebności klas bez wątpienia wpłynęła na ostateczne wyniki. Jak widać na wykresie MDS (rys. 5), najwyraźniej oddzielają się

[26] D.L. Hoover, *Corpus Stylistics, Stylometry, and the Styles of Henry James*, „Style", nr 41 (2) (2007), s. 174–203.

[27] X. Le, I. Lancashire, G. Hirst, R. Jokel, *Longitudinal Detection of Dementia through Lexical and Syntactic Changes in Writing: A Case Study of Three British Novelists*, „Literary and Linguistic Computing", nr 26 (4) (2011), s. 435–461.

najstarsi mówcy, z kolei między pozostałymi zacierają się wyraźne granice, choć widoczne są i skupienia mówców z klas A, B i C, przy czym większe podobieństwo do siebie nawzajem wykazują klasy B i C. Takie wyniki mogą oznaczać, że różnica około dekady to zbyt mało, żeby mówić o wyraźnych zmianach stylu wypowiadania się.

Rys. 5. Różnice między informatorami w wieku 71–80 lat (prefiks A), 82–90 lat (prefiks B), 90–97 lat (prefiks C) a 101–104 lat (prefiks D). Analiza skupień.



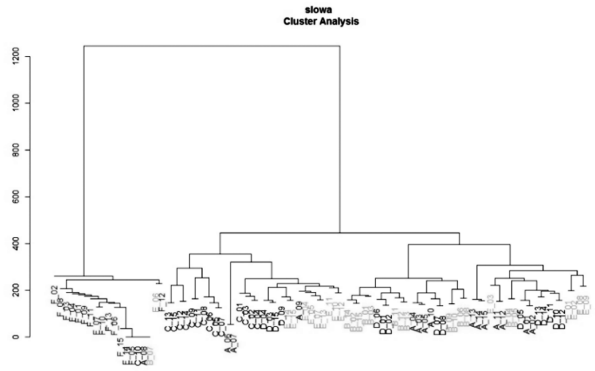
4. Długość wypowiedzi

Korpus czwarty składał się z największej liczby klas (sześciu), ponieważ długości wypowiedzi poszczególnych informatorów były najbardziej zróżnicowane, wahały się od 7 do około 60 tys. słów. Jak już wspomniano, taka rozpiętość wynikała z różnic w czasie trwania wywiadów – występowała (choć z drobnymi wyjątkami[28]) korelacja między czasem trwania wywiadu a liczbą słów – im dłuższy wywiad, tym więcej słów. Wpływ na ilość słów w danej wypowiedzi miały też różnice w sposobie mówienia każdego z informatorów. Klasy wydzielono w odstępach około 10 tys. słów. Podobnie jak w przypadku wieku wyraźnie oddzieliła się najmniej liczna klasa F (7–9 tys. słów – ale tylko 2 mówców), a pozostałe grupują się w mniejsze skupienia

[28] Jako przykład wyjątkowej sytuacji można podać wywiad o największej liczbie słów – 60 435, który trwał ponad dwanaście godzin, podczas gdy drugi z kolei, liczący 59 167 słów – znaczniej krócej, bo niecałe cztery godziny. W tym przypadku o liczbie słów zadecydował sposób mówienia.

(rys. 6). Można wywnioskować, że długość wypowiedzi ma wpływ na klasyfikację, w której wyraźniejszą separację wykazują krótsze teksty (klasy C, E, F), zaś najdłuższe teksty (klasy A i B) należą do najbardziej rozproszonych.

Rys. 6. Różnice między tekstami o długościach: 60k–50k słów (prefiks A), 48k–41k słów (prefiks B), 35k–30k słów (prefiks C), 29k–20k słów (prefiks D), 19k–11k słów (prefiks E) a 9k–7k słów (prefiks F). Analiza skupień.

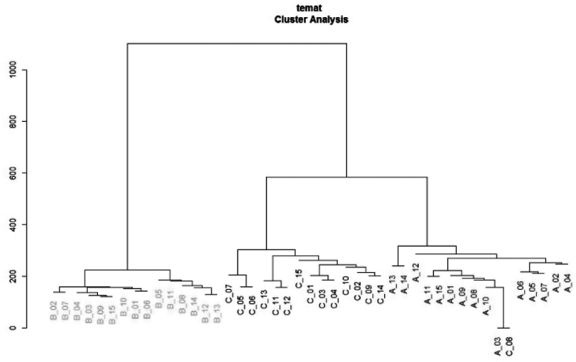


5. Temat

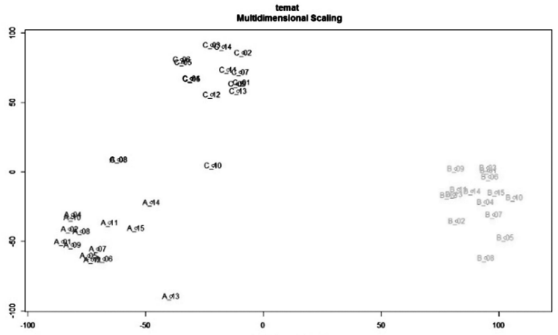
W przypadku korpusu piątego wyniki komputerowej klasyfikacji prawie całkowicie pokryły się z podziałem tematycznym zaprezentowanym w bazie www.relacjebiograficzne.pl – skuteczność wyniosła 98%, a tylko jedna z piętnastu próbek należących do kategorii „ziemiaństwo" została niewłaściwie przyporządkowana do kategorii „Warszawa". Warto zauważyć, że na dendrogramie obie te klasy znalazły się w obrębie jednej, nadrzędnej, gałęzi w opozycji do klasy „wieś", co świadczy o ich większym wzajemnym podobieństwie. Należy także dodać, że poprawnie wyodrębniono trzy klasy pomimo znacznych różnic w ich liczebności[29]. Do grupy najliczniejszej – klasa A; temat „Warszawa" – należało aż 31 mówców, podczas gdy do grupy najmniej licznej – klasa C; temat „ziemiaństwo" – tylko 3 (zob. rys. 7 i 8).

[29] Por.: pkt 3. Wiek i pkt 4. Długość wypowiedzi.

Rys. 7. Różnice między tematami wypowiedzi: „Warszawa" (prefiks A), „wieś" (prefiks B), „ziemiaństwo" (prefiks C). Analiza skupień.



Rys. 8. Różnice między tematami wypowiedzi: „Warszawa" (prefiks A), „wieś" (prefiks B), „ziemiaństwo" (prefiks C). Skalowanie wielowymiarowe.



Podsumowanie

Wszystkie przeprowadzone eksperymenty miały na celu sprawdzenie, czy wiedza *a priori* o pewnych podziałach wewnątrz korpusu (ze względu na płeć, miejsce pochodzenia, wiek, długość wypowiedzi, temat) znajdzie swoje potwierdzenie w ilościowej analizie samego języka. Trzeba bowiem podkreślić, że faktyczna przynależność danych próbek tekstowych do poszczególnych klas nie jest znana komputerowi, który próbuje odnaleźć grupowania tekstów w surowym materiale językowym. To dopiero późniejsze porównanie wyników analizy komputerowej z wiedzą aprioryczną badacza pozwala stwierdzić, w jakim stopniu klasyfikacja spodziewana (idealna) pokrywa się z klasyfikacją rzeczywistą (wynikami eksperymentu).

Zaprezentowane w niniejszym artykule badania mają charakter wstępnej analizy, a przede wszystkim próby na nowej dla stylometrii materii – relacjach biograficznych – jako podstawy do przeprowadzenia badań. Jak wykazano, treści wywiadów biograficznych, a więc teksty języka mówionego, podlegają podobnym prawom, co teksty literackie. Co do skuteczności samej metody należy stwierdzić, że w dwóch przypadkach (płeć autora i temat wypowiedzi) komputer w zasadzie bezbłędnie przeprowadził klasyfikację. W odniesieniu do miejsca urodzenia mówców klasyfikacja rzeczywista nie pokryła się z zakładaną. Z kolei dla wieku i długości wypowiedzi wyniki okazały się podobne: najmniejsze klasy zostały prawidłowo rozróżnione, podczas gdy pozostałe grupowały się w niejednorodnych skupieniach. Powyższe wnioski dowodzą, że nawet na podstawie tak prostych eksperymentów stylometrycznych można wykazać pewną skuteczność metod ilościowych w badaniu cech różnicujących teksty języka mówionego.

Bibliografia

Bartmiński J., *O wartościach słowa mówionego*, [w:] *Historia mówiona w świetle etnolingwistyki*, red. S. Niebrzegowska-Bartmińska, S. Wasiuta, Lublin 2008, s. 9–16.

Burrows J.F., *„Delta": A Measure of Stylistic Difference and a Guide to Likely Authorship*, „Literary and Linguistic Computing", nr 17 (2002), s. 267–287.

Burrows J.F., *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, Oxford 1987.

Eder M., *Metody ścisłe w językoznawstwie i pułapki pozornego obiektywizmu. Przykład stylometrii*, „Teksty Drugie", nr 2 (2014), s. 90–105.

Eder M., *Style-markers in authorship attribution: a cross-language study of the authorial fingerprint*, „Studies in Polish Linguistics", nr 6 (2011), s. 99–114.

Eder M., Kestemont, M., Rybicki J., *Stylometry with R: A package for computational text analysis,* „R Journal", nr 16 (1) (2016), s. 107–121.

Eder M., Rybicki J., Młynarczyk K. [et al.], *1000 Novels Corpus*, CLARIN-PL digital repository, 2016.

Gocół D., *Opozycja swoi/obcy w relacjach radomskiego Czerwca '76*, [w:] *Tekst – gatunek – dyskurs na przełomie XX i XXI wieku*, red. J. Szadura, Lublin 2012, s. 135–152.

Grochola-Szczepanek H., *Badania fokusowe mowy mieszkańców wsi*, „Socjolingwistyka" nr 20, (2006), s. 19–35.

Herz J., Bellaachia A., *The authorship of audacity: Data mining and stylometric analysis of Barack Obama speeches*, [w:] *Proceedings of the International Conference on Data Mining*, b.m. 2014.

Hoover D.L., *Corpus Stylistics, Stylometry, and the Styles of Henry James*, „Style", nr 41 (2) (2007), s. 174–203.

Labocha J., *Pragmatyczne mechanizmy składni języka mówionego*, „Slavia Occidentalis", nr 69 (2012), s. 139–145.

Le X., Lancashire I., Hirst G., Jokel R., *Longitudinal Detection of Dementia through Lexical and Syntactic Changes in Writing: A Case Study of Three British Novelists*, „Literary and Linguistic Computing", nr 26 (4) (2011), s. 435–461.

Lewandowska I., *Wywiad jako technika zdobywania informacji źródłowych w badaniu historii najnowszej*, „Echa Przeszłości", t. V (2004), s. 279–299.

Mandravickaitė J., Krilavičius T., *Stylometric Analysis of Parliamentary Speeches: Gender Dimension*, [w:] *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, red. T. Erjavec [et al.], Valencia 2017, s. 102–107.

Pacławska E., *Zróżnicowanie gatunków mowy w tekstach historii mówionej*, [w:] *Historia mówiona w świetle etnolingwistyki*, red. S. Niebrzegowska-Bartmińska, S. Wasiuta, Lublin 2008, s. 47–62.

Pęzik P., *Język mówiony w NKJP*, [w:] *Narodowy Korpus Języka Polskiego*, red. A. Przepiórkowski, M. Bańko, R. Górski, B. Lewandowska-Tomaszczyk, Warszawa 2012, s. 37–48.

Rybicki J., *Vive la différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies*, „Digital Scholarship in the Humanities", nr 31 (4) (2016), s. 746–761.

Weidman S.G., O'Sullivan J., *The limits of distinctive words: Re-evaluating literature's gender marker debate*, „Digital Scholarship in the Humanities", nr 33 (2) (2017), s. 374–390.

Magdalena Kądzioła

Factors Differentiating the Statements of Narrators: A Stylometric Analysis of Biographical Interviews

This article proposes ways to analyse the content and metadata of biographical interviews using statistical methods. The basis for this series of stylometric experiments was a specially created corpus exceeding 1.2 million lexical units in size and composed of texts extracted from selected biographical interviews from the Oral History Archive, the History Meeting House, and the KARTA Centre available on the website www.relacjebiograficzne.pl. Research was based on the content of biographical interviews with forty-one people assigned to three thematic categories: 'Warsaw,' 'the village,' and 'gentry.' The main goal of the experiments was to determine which linguistic factors differentiate speakers and which features (gender, place of origin, age, length of speech, or topic) can influence this classification. This research was carried out using quantitative linguistics methods, and the conclusions we have arrived at allow for the determination of the direction of further work in the field of the stylometry of spoken language.

Keywords: **biographical accounts, stylometry, spoken language, classification of texts, metadata**