

Maciej Eder

(Instytut Języka Polskiego PAN, Uniwersytet Pedagogiczny)

SŁOWA ZNACZĄCE, SŁOWA KLUCZOWE, SŁOWOZBIORY – O STATYSTYCZNYCH METODACH WYSZUKIWANIA WYRAZÓW ISTOTNYCH*

Pojęcie słów-kluczy bywa używane w kilku różnych kontekstach. Z jednej strony mogą to być pewnego rodzaju słowa-hasła czy też „słowa sztandarowe”¹, a zatem słowa (oraz wyrażenia), które można wypisać na sztandarach i transparentach. Z drugiej strony o słowach-kluczach mówi się w kontekście metadanych bibliologicznych, tak ważnych w wyszukiwarkach internetowych czy w szeroko rozumianej inżynierii informacji². Słowa-klucze to wreszcie – w ujęciu Wierzbickiej – pewien zasób kilkudziesięciu pojęć podstawowych, uniwersalnych w tym znaczeniu, że nie zależą one od żadnego języka czy kręgu kulturowego³.

Niniejszy artykuł również będzie poświęcony wyrazom istotnym, przy czym owa istotność będzie rozumiana jako wynik procedury empirycznej, opartej na różnego rodzaju miarach statystycznych. Zostaną pokrótce przedstawione trzy metody, których istotą jest wydobywanie z korpusu (zbioru tekstów) słów wyróżniających się pod względem frekwencji, względnie słów, które w sposób nieprzypadkowy występują obok siebie. Po pierwsze będzie to metoda słów kluczowych (*keywords analysis*), po drugie metoda Zeta Burrowsa–Craiga, po trzecie zaś zyskująca ostatnio ogromną popularność metoda modelowania tematycznego (*topic modeling*). Pomysł zastosowania miar statystycznych do wyszukiwania w tekstach słów bądź fraz istotnych (np. do automatycznego przeszukiwania dużych baz danych) nie jest oczywiście nowy; pojęcie słowa kluczowego, obok pojęcia kolokacji, należy do podstawowego repertuaru metod językoznawstwa kwantytatywnego niemal od początku istnienia tej dyscypliny⁴.

Założenia teoretyczne, do których pośrednio lub bezpośrednio odwołują się wszystkie próby wydobywania z korpusu jakichś elementów nacechowanych (już to leksykalnych, już to składniowych), są w zasadzie zawsze takie same. Po pierwsze, wolno założyć, że dystrybucja poszczególnych słów w korpusie (oraz *ex fortiori*

* Praca niniejsza została wykonana w ramach projektu UMO-2013/11/B/HS2/02795 finansowanego ze źródeł Narodowego Centrum Nauki.

¹ Walery Pisarek, *Polskie słowa sztandarowe i ich publiczność*, Universitas, Kraków 2002.

² Wiesław Babik, *Słowa kluczowe*, Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków 2010.

³ Anna Wierzbicka, *Understanding cultures through their key words*, Oxford University Press, New York 1997.

⁴ Pierre Guiraud, *Les caractères statistiques du vocabulaire*, PUF, Paris 1954; Michael Oakes, *Statistics for corpus linguistics*, Edinburgh University Press, Edinburgh 1998; Adam Kilgariff, *Language is never ever ever random*, „Corpus Linguistics and Linguistic Theory” 2005, nr 1/2, s. 263–276.

w języku) jest zmienną losową o dającej się zmierzyć charakterystyce. Nie oznacza to oczywiście, że użyciem słów rządzi przypadek; przeciwnie, poszczególne elementy są względem siebie do pewnego stopnia zależne, np. obok przyimka najczęściej wystąpi rzeczownik bądź zaimek w ściśle określonym przypadku. Użytkownik języka ma jednak ogromną swobodę wyboru słownictwa, co sprawia, że dystrybucja słów w każdym tekście podlega fluktuacjom.

Po drugie, dla całego korpusu można oszacować średnie frekwencje wszystkich słów czy fraz, a zatem otrzymać jakieś przybliżenie do *normy* (rozumianej statystycznie), czyli przeciętnego użycia poszczególnych elementów języka. Intuicja podpowiada, że wyrazy bardzo istotne dla prozy danego pisarza (gatunku, stylu, rejestru) będą się charakteryzowały wyraźnie częstszym użyciem niż wskazywałaby owa „norma”. Na tej właśnie intuicji opiera się następne założenie, a można je streścić następująco: przez porównanie frekwencji wyrazów obliczonych dla całego korpusu z ich frekwencjami w badanej próbce można oszacować, czy zauważone różnice między „normą” (wartością oczekiwaną) i badaną próbką (wartością obserwowaną) mogły być dziełem przypadku, czy też są to różnice znaczące. Oszacowanie statystycznej istotności owej różnicy odbywa się za pomocą jednego ze standardowych testów statystycznych⁵.

Powyższa ogólna zasada znajdowania zjawisk nietypowych w korpusie stosuje się zarówno do metod przedstawionych w dalszej części artykułu, jak i, *mutatis mutandis*, do bardzo wielu innych procedur językoznawstwa kwantytatywnego oraz korpusowego.

SŁOWA KLUCZOWE

Podstawową, a zarazem najprostszą metodą znajdowania słów ważnych jest wspomniana powyżej analiza słów kluczowych (*keywords analysis*). Najbardziej oczywiste jej zastosowanie to wydobycie słów charakteryzujących pisarstwo jakiegoś autora⁶, ale można sobie wyobrazić wiele innych pytań badawczych, np. porównanie tekstów napisanych przez kobiety i przez mężczyzn pod kątem słownictwa różnicującego obie grupy⁷.

W poniższym przykładzie przeprowadzona zostanie analiza kontrastywna między dwoma odmianami polszczyzny – literacką i popularno-publicystyczną. W tym celu zebrane zostały dwa podkorpusy: na jeden z nich składa się 100 powieści polskich obejmujących teksty 34 autorów z końca XIX i początku XX wieku (Sienkiewicz, Berent, Dąbrowska, Dmochowska, Reymont, Żeromski, Zapolska etc.)⁸, drugi zaś jest kompletnym zbiorem artykułów opublikowanych w serwisie Onet.pl w przeciągu jednego pełnego miesiąca, mianowicie września 2015 roku⁹.

⁵ Adam Kilgarriff, *Comparing corpora*, „International Journal of Corpus Linguistics” 2001, nr 1, s. 1–37.

⁶ Por. Jadwiga Sambor, *Badania statystyczne nad słownictwem. Na materiale „Pana Tadeusza”*, Ossolineum, Wrocław 1969.

⁷ James Pennebaker, *The secret life of pronouns: What our words say about us*, Bloomsbury, New York 2011, s. 39–72.

⁸ Cały zbiór dostępny jest pod adresem: https://github.com/computationalstylistics/100_polish_novels [dostęp 15.04.2016]. Wyboru tekstów do korpusu dokonał Jan Rybicki.

⁹ Zbiór ten został sporządzony na potrzeby Maratonu Analizy Danych zorganizowanego przez portal Onet.pl oraz Społeczność Entuzjastów Języka R „eRka”.

Mimo ogromnej dysproporcji w liczbie próbek – 100 powieści *contra* 47 540 artykułów – oraz jeszcze większej różnicy w długości próbek – dziesiątki tysięcy słów w jednej powieści *contra* średnio kilkadziesiąt słów w przeciętnym artykule – oba podkorpusy są porównywalne jako całość, liczą bowiem odpowiednio 9 oraz 12 milionów wyrazów. Zaznaczyć należy, że oba podkorpusy zostały wykorzystane w postaci nielematyzowanej. Oznacza to, że form wyrazów pojawiających się w poszczególnych tekstach nie sprowadzano do formy podstawowej (tj. słownikowej), lecz poddawano analizie w postaci oryginalnej. Pojawiające się w korpusie różne formy fleksyjne, np. „byliśmy”, „są”, „jestem”, „być”, traktowano zatem jako niezależne jednostki leksykalne i obliczano frekwencje dla każdej z nich z osobna. Mimo pewnej ułomności metodologicznej takiego ujęcia zyskuje się w ten sposób wgląd nie tylko w dystrybucję poszczególnych słów, ale i w ich rzeczywiste użycie¹⁰.

Procedura badawcza polega na obliczeniu częstości występowania każdego słowa w każdym tekście i sporządzeniu ogólnej, uśrednionej listy frekwencyjnej z jednej strony dla podkorpusu literackiego, z drugiej zaś strony dla podkorpusu popularno-publicystycznego. Następnie frekwencja każdego słowa zostaje porównana na obu listach i dla każdej zauważonej różnicy szacowana jest jej statystyczna istotność. Efektem owego stosunkowo prostego porównania jest miara „kluczowości”, która jest tym większa, im bardziej dana wartość obserwowana odstaje od założonej wartości teoretycznej. Przy obliczaniu „kluczowości” stosuje się albo test chi-kwadrat, albo test log-likelihood: powszechnie znane i szeroko stosowane testy na istotność statystyczną.

Początek listy słów charakterystycznych dla polszczyzny literackiej przełomu XIX i XX wieku został przedstawiony w poniższej tabeli. Do obliczania miary „kluczowości” zastosowano program AntConc, dostępny na darmowej licencji¹¹.

Ranga	Frekwencja	„Kluczowość”	Słowo
1	288 314	47 841,008	i
2	22 706	28 119,910	pani
3	26 170	27 853,178	ja
4	25 369	27 134,097	pan
5	231 849	22 614,184	się
6	26 281	17 682,414	mu
7	9 820	17 151,597	rzekł
8	174 008	16 591,529	nie
9	34 711	16 392,381	jej
10	41 450	14 790,948	tak
...

¹⁰ Zainteresowani tematyką wpływu lematyzacji na właściwości statystyczne tekstów w języku polskim garść danych empirycznych znajdują w studium: Rafał L. Górski, Maciej Eder, *Stylistic fingerprints, POS tags and inflected languages: a case study in Polish*, w druku.

¹¹ Program jest dostępny na stronie: <http://www.laurenceanthony.net/software/antconc> [dostęp 15.04.2016].

W przeciwieństwie do tradycyjnie rozumianych słów-kluczy, zamykających się zwykle liczbą kilkudziesięciu jednostek leksykalnych, powyższa lista wygenerowana za pomocą algorytmu nie ma jasno określonego końca: poszczególne słowa ułożone są w kolejności od najbardziej do najmniej „kluczowego”. W praktyce jednak silną miarę „kluczowości” wykazuje zaledwie niewielka część słów, za którymi idzie bardzo długi pochód wyrazów o „kluczowości” bliskiej zera. Lista jest zresztą symetryczna, ponieważ po słowach kluczowych dla podzbioru pierwszego umieszczone zostają słowa coraz bardziej nietypowe, czyli, inaczej rzecz ujmując, słowa kluczowe dla drugiego podzbioru. Oto 50 słów najmocniej wyróżniających literaturę powieściową na tle piśmiennictwa prasowego:

i, pani, ja, pan, się, mu, rzekł, nie, jej, tak, go, mnie, a, mi, oczy, ją, ty, jakby, tu, co, ku, cóż, pana, niech, panie, on, rzekła, panna, znowu, nią, ani, lecz, nic, ona, ale, no, zaraz, zawołał, niej, sobie, chwili, ręce, twarz, nagle, nim, tam, głowę, zaś, rękę, człowiek, ...

Z kolei dla tekstów popularno-publicystycznych najbardziej charakterystyczne są następujące wyrazy:

w, i, roku, oraz, zł, proc, jest, polski, m, tys, in, r, są, podczas, mecz, będzie, ponad, będą, sezonie, Polsce, również, reprezentacji, wcześniej, osób, Europy, także, września, lat, klubu, Polska, ligi, sierpnia, uchodźców, km, został, zespół, czyli, gry, drużyny, przypadku, zawodników, premier, zespołu, sytuacji, np, prezydent, pl, obecnie, mogą, według, ...

Krótkiego komentarza wymagają oba zestawy słów, już na pierwszy bowiem rzut oka widać nie tylko skądinąd oczywiste różnice tematyczne, ale również wyrazy niezwiązane bezpośrednio z treścią.

Niemal na samym początku listy wyrazów popularno-publicystycznych pojawia się tajemnicze słowo „l”, którego silną pozycję należy wyjaśnić charakterystycznym dla tabloidów zwyczajem podawania wieku omawianych osób, por. „Kamil Stoch (28 l.) wygrał trzy ostatnie konkursy indywidualne”¹². Z kolei „r” jest powszechnie używanym skrótem używanym w datach: „Trybunał miał badać pytanie prawne Sądu Okręgowego w Lublinie z 2014 r. o zgodność z konstytucją przepisu ustawy z 1991 r. o związkach zawodowych”¹³. Na dalszych pozycjach listy uwagę zwracają wyrazy (zbitki liter?), „m” oraz „in”, które w istocie przynależą do jednego skrótu „m.in.”, tak częstego w popularnej publicystyce. Trudno również nie zauważyć innych skrótów, takich jak „zł”, „proc.”, „tys.”, „np.”, itd., rzeczą jednak w szczególności sposób rzucającą się w oko jest nadreprezentacja zaimków oraz partykuł po stronie literatury XIX-wiecznej oraz słownictwo tematycznie związane ze sportem i trochę mniej wyraźnie z polityką – po stronie prasy XXI wieku.

Uważna lektura obu list słów zdradza jeszcze jedną osobliwość, mianowicie wysoką i trudną do wyjaśnienia wagę przyimków w podkorpusie prasowym: „w”, „podczas”, „ponad”, „według”, oraz istotną rolę spójników w zbiorze powieści: „i”, „a”, „ani”, „lecz”, „ale”, zapewne dającą się wyjaśnić bogatszą składnią, w tym hipotaksą charakteryzującą literaturę piękną. Znacznie trudniej natomiast wyjaśnić nadobecność zaimka zwrotnego „się” w podkorpusie powieściowym. Zapewne należy wiązać ten fakt z odmienną dystrybucją czasowników zwrotnych w obu

¹² Artykuł przywołany za dziennikiem „Fakt”, plik 4426015965.txt

¹³ Artykuł przywołany za Polską Agencją Prasową, plik 4400597564.txt

podzbiorach, ale spodziewalibyśmy się wyniku dokładnie odwrotnego, tzn. większej frekwencji zaimka „się” w prasie, a nie w literaturze.

SŁOWA O REGULARNEJ DYSTRYBUCJI

Przedstawiona powyżej metoda ekstrakcji słów statystycznie istotnych pozwalała na sformułowanie ciekawych spostrzeżeń na temat składni oraz dystrybucji zaimków w obu podkorpusach – co samo w sobie ma sporą wartość poznawczą – trudno jednak zaprzeczyć, że wyłonione przez algorytm wyrazy synsemantyczne są jednak nieprzekonujące w roli słów-kluczy. Spodziewalibyśmy się raczej czegoś w rodzaju szybkiego wglądu w charakterystyczne słownictwo literatury pięknej z jednej strony i słownictwo tabloidów z drugiej, a nie listy słów synsemantycznych.

Częściowe rozwiązanie przynieść może zastosowanie metody wprowadzonej przez Burrowsa i przeznaczonej do badania autorstwa tekstów anonimowych¹⁴. Metoda owa, nazwana przez Burrowsa określeniem „Zeta”, rozszerzona następnie przez Craiga w jego rozprawie na temat kanonu dzieł Szekspira¹⁵, była z powodzeniem stosowana nie tylko w atrybucji autorskiej, ale i w szerzej rozumianych badaniach literaturoznawczych¹⁶. Założenia metodologiczne tej techniki są bardzo podobne do przedstawionej powyżej analizy słów kluczowych, przy jednej istotnej różnicy. Otóż zamiast mierzyć frekwencję poszczególnych słów w korpusie, Zeta najpierw dzieli teksty z korpusu na niezbyt duże próbki (segmenty), a następnie bada, w ilu segmentach dane słowo wystąpiło – nie bezpośrednio tedy frekwencja słowa, lecz frekwencja segmentów stanowi podstawę dalszych porównań. W konsekwencji słowa bardzo częste (a więc słowa synsemantyczne) są z definicji filtrowane, dlatego lista słów znaczących jest znacznie bardziej zogniskowana na treści.

W części eksperymentalnej został wykorzystany ten sam korpus co w poprzednim teście, czyli zestaw 100 powieści oraz 47 540 artykułów prasowych. Do obliczeń został użyty darmowy pakiet Stylo przeznaczony dla środowiska R¹⁷. Wyłonione przez algorytm wyrazy różnią się od omawianych powyżej słów kluczowych: w szczególności rzuca się w oczy brak wielu słów synsemantycznych. Następujące wyrazy zostały uznane za istotne dla zbioru literackiego:

oczy, rzekł, jakby, cóż, ku, ty, niech, ręce, głowę, twarz, rękę, głosem, znowu, zaraz, nagle, pana, serce, pan, człowiek, twarzy, zawołał, panie, swego, drzwi, duszy, Bóg, głos, myśl, myśli, oczach,

¹⁴ John Burrows, *All the way through: testing for authorship in different frequency strata*, „Literary and Linguistic Computing” 2007, nr 22, s. 27–48.

¹⁵ Hugh Craig, Arthur F. Kinney, *Shakespeare, Computers, and the Mystery of Authorship*, Cambridge University Press, Cambridge 2009.

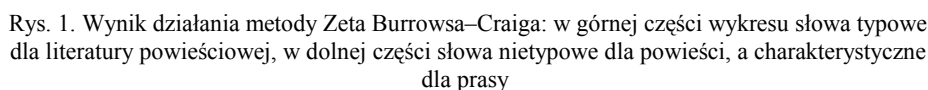
¹⁶ David Hoover, *Teasing out authorship and style with t-tests and Zeta*, „Digital Humanities 2010: Conference Abstracts”, King’s College, London 2010, s. 168–170; Alexis Antonia, Hugh Craig, Jack Elliott, *Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution*, „Literary and Linguistic Computing” 2014, nr 29 (2), s. 147–163; Jan Rybicki, *Vive la différence: tracing the (authorial) gender signal by multivariate analysis of word frequencies*, „Digital Scholarship in the Humanities”, w druku.

¹⁷ Maciej Eder, Mike Kestemont, Jan Rybicki, *Stylometry with R: A package for computational text analysis*, „R Journal” 2016, nr 16, <https://journal.r-project.org/archive/accepted/>.

Słowa istotne dla podkorpusu popularno-publicystycznego są następujące:

l, oraz, meczu, polski, m, in, zł, tys, mecz, wcześniej, trener, Polsce, ponad, sezonie, roku, podczas, r, proc, Europy, czyli, reprezentacji, września, klubu, mln, sierpnia, sytuacji, ok, euro, obecnie, również, Polska, dzięki, ligi, według, zespół, osób, drużyny, gry, zespołu, piłkarz, zdaniem, przypadku, kolejne, spotkania, zawodników, związku, został, ostatnio, latach, spotkaniu, ...

Program Stylo umożliwia graficzne zobrazowanie istotności słów odnalezionych przez metodę Zeta. Na rys. 1 wyniki zostały pokazane na wykresie: im bardziej dane słowo jest oddalone od linii bazowej, tym wyższa jest jego „kluczowość”. Słowa typowe dla literatury znajdują się w górnej części wykresu, słowa charakterystyczne dla prasy – w dolnej części.



Wyniki pokazują dość jednoznacznie dominację tematyki sportowej w prasie; słownictwo literatury pięknej jest nieco bardziej rozmyte, ale i tu widać pewne kręgi tematyczne, w tym określenia części ciała („oczy”, „ręce”, „głowę”, „twarz”, „rękę”, „usta”) czy *verba dicendi* („rzekł”, „zawołał”, „rzekła”, „zdawało”, „widział”).

SŁOWA POWIĄZANE TEMATYCZNIE

Niezwykle ciekawą i zyskującą sobie coraz większą popularność metodą analizy dużych zbiorów danych jest tzw. modelowanie tematyczne (*topic modeling*), oparte na bardzo skomplikowanym algorytmie wyszukiwania wyrazów wykazujących tendencję do współwystępowania w niedalekim sąsiedztwie¹⁸.

Zasadę działania metody łatwiej zrozumieć przez przywołanie pojęcia kolokacji – jednego z fundamentów językoznawstwa korpusowego. Otóż idea kolokacji w pewnym uproszczeniu zasadza się na tym, że jeśli słowo *A* w korpusie występuje z jakimś prawdopodobieństwem $P(A)$, słowo zaś *B* z prawdopodobieństwem $P(B)$, to prawdopodobieństwo współwystąpienia obu tych słów w bezpośrednim sąsiedztwie wynosi $P(A \cap B) = P(A) \times P(B)$, czyli jest niezwykle niskie. Tymczasem w języku pojawia się pewna liczba połączeń wyrazowych, których frekwencja rzeczywista jest wielokrotnie wyższa, niż wynikałoby z teoretycznego prawdopodobieństwa (np. *mocna + kawa* lub *pies + ogrodnika*). Takie właśnie połączenia zwykło się nazywać kolokacjami, a podstawą ich wyodrębniania jest wyłącznie miara statystyczna orzekająca o tym, jak bardzo ich rzeczywiste występowanie różni się od frekwencji spodziewanej.

Ogólna idea stojąca za modelowaniem tematycznym jest w gruncie rzeczy bardzo podobna, lecz stopień komplikacji problemu jest bez porównania większy. Celem analizy jest bowiem nie tyle szukanie par wyrazów, ile wyodrębnianie całych konstelacji „lubiących się” słów. Zamiast tedy wyszukiwać pewną liczbę par słownych (kolokacji), wyszukuje się całe „tematy”, obejmujące wiele współwystępujących słów jednocześnie. Modelowanie tematyczne było w zamyśle przeznaczone do przeszukiwania dużych zbiorów danych tekstowych; celem miało być wyszukiwanie treści (w tym: sporządzanie automatycznych streszczeń) w kolekcjach dokumentów, porządkowanie dokumentów pod względem ich podobieństwa tematycznego, np. w bibliotekach cyfrowych czy w centrach informacji naukowej. W ostatnich latach metoda zdobywa jednak coraz większą popularność w wielkoskalowych badaniach literaturoznawczych¹⁹, historycznych²⁰ czy językoznawczych²¹.

¹⁸ David Blei, *Probabilistic topic models*, „Communications of the ACM” 2012, nr 55/4, s. 77–84.

¹⁹ Matthew Jockers, *Macroanalysis: Digital Methods and Literary History*, University of Illinois Press, Urbana 2013; Christof Schöch, *Topic modeling genre: an exploration of French classical and enlightenment drama*, „Digital Humanities Quarterly”, w druku.

²⁰ Benjamin M. Schmidt, *Words alone: dismantling topic models in the humanities*, „Journal of Digital Humanities” 2012, nr 1, <http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/> [dostęp 15.04.2016].

²¹ Urszula Modrzyk, *Topic modeling i językoznawstwo? Analiza semantyki dwóch przymików złożonych*, w druku.

Najpopularniejszą implementacją modelowania tematycznego jest algorytm znany pod nazwą LDA (*Latent Dirichlet Allocation*)²². Od strony matematycznej metoda jest bardzo złożona i wymaga dużych zasobów komputera do przeprowadzenia obliczeń; metoda zresztą działa najlepiej wtedy, gdy zbiór analizowanych tekstów także jest duży. Działanie algorytmu opiera się na założeniu, że słowa powiązane tematycznie będą miały skłonność do współwystępowania w jednym dokumencie i będą bardzo słabo reprezentowane w innych dokumentach. Na przykład w powieści detektywistycznej można się spodziewać słów typu „poszlaka”, „ślady”, „trup”, „policja” itp., a ich obecność (i współobecność) będzie tym bardziej znacząca, im rzadziej wystąpią one w pozostałych tekstach korpusu, dajmy na to, w powieści historycznej czy sentymentalnej. LDA jest matematycznym modelem służącym do wydobywania owych ukrytych relacji między słowami, a dobieranie parametrów tego modelu odbywa się w ten sposób, że z rzeczywistego zbioru danych tekstowych zostaje wyłoniony – w sposób przypominający praktyki dadaistów – bezsensowny korpus czysto losowych „tekstów”. Porównanie próbek rzeczywistych z losowymi ujawnia nieprzypadkowe połączenia wyrazowe w próbkach rzeczywistych: im większa różnica danych obserwowanych i danych teoretycznych (tj. losowych), tym większy jest związek badanych wyrazów. Cała procedura zostaje następnie powtórzona w bardzo dużej liczbie iteracji, dzięki czemu niedoskonały z początku model sukcesywnie się optymalizuje przez eliminację współwystąpień przypadkowych. Czułość metody LDA wzrasta znacząco, jeśli wcześniej usunąć z korpusu bardzo częste słowa synsemantyczne: „się”, „w”, „i”, „lub”, „oraz” itd. (partykuły, przyimki, spójniki, większość zaimków)²³.

Efektom działania algorytmu są dwa zbiory liczb zorganizowane w postaci dwóch macierzy: jedna macierz zawiera prawdopodobieństwa wystąpienia kolejnych „tematów” w poszczególnych dokumentach, druga natomiast macierz zawiera słowa, które wystąpiły w danym „temacie”. Należy w tym miejscu z całą stanowczością zaznaczyć, że omawiana procedura opiera się tylko i wyłącznie na frekwencji poszczególnych wyrazów, bez żadnej dodatkowej informacji na temat ich powiązań semantycznych. Działanie algorytmu bywa zdumiewająco trafne, zgodne z intuicją użytkownika języka (wyłoniłoby się wyrazy podobne znaczeniowo), wynika to jednak najwyraźniej z faktu, że relacje semantyczne dają się w jakimś stopniu wy-modelować na podstawie niewidocznych gołym okiem podobieństw frekwencyjnych.

Mimo że algorytm zdaje się rozumieć znaczenia wyrazów, nie wszystkie odnalezione w ten sposób „tematy” (*topics*) wykazują rzeczywistą spójność treściową czy tematyczną. Często obok „tematów” zawierających np. grupę powiązanych znaczeniowo rzeczowników pojawiają się np. zbiory słów zawierające *verba dicendi*, „tematy” matajęzykowe, wreszcie zestawy słów z pozoru niepowiązanych. Nazywanie ich wszystkich „tematami” nie do końca oddaje istotę rzeczy. Dlatego może lepiej mówić tutaj o konstelacjach wyrazów „lubiących się” czy też o słowozbiorach²⁴ niż o tematach w utartym znaczeniu tego słowa.

²² David Blei, Andrew Ng, Michael Jordan, *Latent Dirichlet Allocation*, „Journal of Machine Learning Research” 2003, nr 4–5, s. 993–1022.

²³ Lisa Rhody, *The story of stopwords: topic modeling an ekphrastic tradition*, „Digital Humanities 2014: Book of Abstracts”, Lausanne 2014, s. 328–330.

²⁴ Termin „słowozbiór” w znaczeniu współwystępujących słów wyznaczonych przez algorytm LDA został ukuty na seminarium z językoznawstwa kwantytatywnego prowadzonym w Pracowni Metodologicznej IJP PAN. Autorami terminu są Rafał L. Górski, Urszula Modrzyk i piszący te słowa.

Przejdźmy do części eksperymentalnej. Testy metody LDA przeprowadzono na tym samym co uprzednio korpusie 100 powieści polskich z przełomu XIX i XX wieku. Do obliczenia frekwencji słów w korpusie użyty został ponownie pakiet Stylo, samo zaś modelowanie tematyczne wykonano za pomocą pakietu Mallet, również dostępnego w środowisku R. Przeprowadzona procedura badawcza wyglądała następująco. Po pierwsze, ze wszystkich tekstów zostały usunięte zaimki, przymyki, partykuły, spójniki i niektóre przysłówki (stoplista zawierała 326 wyrazów). Następnie usunięte zostały wszystkie słowa, które występowały w bardzo małej liczbie tekstów, nawet jeśli same miały wysoką frekwencję (ustawiono próg obecności w 10% tekstów, a więc w przynajmniej 10 powieściach ze 100). W ten sposób usunięto większość nazw własnych, w tym wiele imion bohaterów powieści. Tak przygotowany korpus podzielono automatycznie na segmenty (próbki) o długości 1000 wyrazów. Liczba próbek oczywiście zależała od długości poszczególnych powieści; cały korpus liczył 8059 próbek. Następnym etapem było uruchomienie algorytmu LDA. Parametrami wejściowymi użytymi do obliczeń było 50 „tematów” (słowozbiorów) wyłonionych w 500 iteracjach.

Niektóre z uzyskanych w ten sposób słowozbiorów bez większych kłopotów daje się objąć jakąś zbiorczą nazwą, np. słowozbiór 1 można by roboczo zatytułować *Wsi spokojna, wsi wesola*:

konie, dzieci, roboty, wsi, dzień, pola, lasu, drodze, ziemi, chłop, wóz, koni, drogi, ziemia, polu, dwór, spod, wieś, pole, mieli, dworze, tyle, konia, słońce, dworu, polach, śniegu, stajni, ganku, drzewa, dni, śnieg, lesie, dziedzica, las, robotę, zboża, człowiek, wrócił, chłopci, dziedzic, ganek, wiatr, chłopów, gęsi, chleb, parę, robota, ludźmi, mleka, ...

Słowozbiór 35 można by nazwać *W kościelnej kruchcie*:

ksiądz, kościoła, Bóg, księdza, Boga, kościół, proboszcz, kościele, Szymon, klasztoru, Bogu, ołtarza, matki, ks, świętego, modlitwy, święty, Boże, rzekł, wiary, proboszcza, księżę, księdzem, ojcie, świętej, świętych, kaplicy, księdzu, św, mszy, klasztor, klasztorze, księży, bożej, modlił, krzyż, Piotr, matka, świętym, boskiej, modlić, nabożeństwo, dzieci, ojciec, chwila, mury, Paweł, święte, śmierci, Chrystus, ...

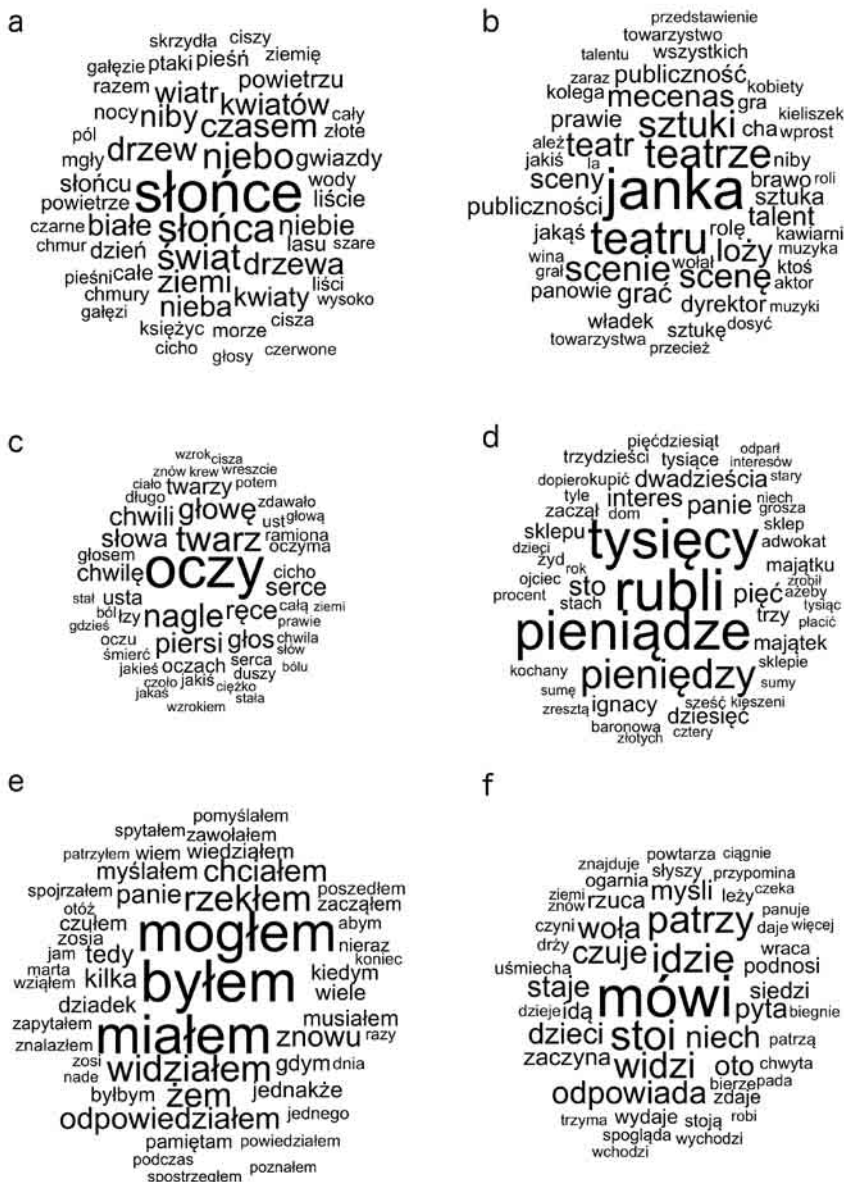
Z kolei słowozbiór 20 dałoby się zamknąć określeniem *Gimnazjum*:

klasy, szkoły, klasie, profesor, gimnazjum, lekcji, nauczyciel, dyrektor, szkole, wreszcie, lekcje, dzieci, klasa, książki, kolega, szkoła, między, nauki, kolegów, przecie, uczyć, dopiero, pierwszej, uczniów, toteż, chłopcy, itd, wszyscy, nauczyciela, uczeń, sali, całej, czytać, Stefan, oczyma, pewnego, profesora, uczył, dwu, siódmej, ażeby, ławki, Jędrak, ogóle, tedy, pewnej, Marcin, podczas, koledzy, miejsca, ...

Najbardziej zdumiewa oczywiście fakt, że automatycznie wyłonione szeregi wyrazów układają się w jednolite tematycznie grupy, widoczne nawet gołym okiem. Powtórzmy: algorytm nie ma żadnej wiedzy o rzeczywistych relacjach semantycznych. Komputer pozbawiony jest intuicji – tak oczywistej dla użytkownika języka – że słowa „lekcje”, „książki”, „szkoła”, „profesor” itp. odwołują się do tej samej rzeczywistości pozatekstowej, a mimo to wszystkie te wyrazy skutecznie włącza do jednego słowozbioru.

Wyniki modelowania tematycznego są jeszcze bardziej przekonujące, jeśli przedstawić je w postaci graficznej. Jako że udział poszczególnych słów w danym słowozbiorze jest zróżnicowany, dość wygodną metodą ich wizualizacji jest

chmura słów²⁵, tj. takie ułożenie słów na wykresie, żeby ich wielkość odpowiadała frekwencji i żeby najistotniejsze wyrazy dla danego słowozbioru znalazły się w środku wykresu. Kilka przykładowych chmur słów zostało przedstawionych na rys. 2a–f, m.in. słowozbiór 26 *Przyroda* (rys. 2a), zawierający takie słowa, jak

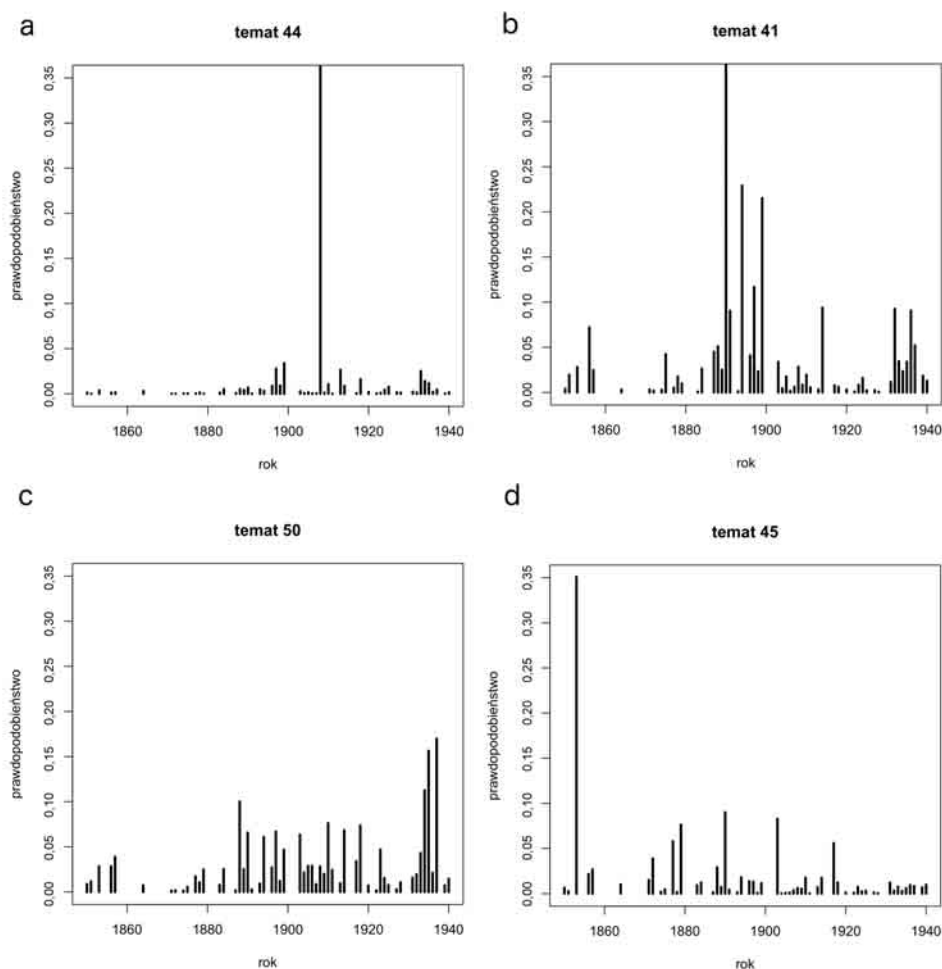


Rys. 2. Sześć przykładowych słowozbiorów z korpusu 100 powieści polskich, przedstawionych w postaci chmur słów

²⁵ Matthew Jockers, *Macroanalysis...*, s. 125.

„słońce”, „ziemia”, „las”, „liście”, „gwiazdy”; słowozbiór 8 *Teatr* (rys. 2b), w skład którego wchodzi kilka imion oraz słowa typu „scena”, „publiczność”, „brawo”, „sztuka” czy „talent”; wreszcie słowozbiór 41 *Pieniądz i kapitał* (rys. 2d), na który składają się wyrazy „tysiące”, „pieniędzy”, „rubli”, „interes”, „sklepu”, „Żyd”, „majątek” etc. Wizualizacja za pomocą chmury słów nie tylko daje szybki wgląd w zawartość danego słowozbioru, ale także pokazuje istotność obecnych w nim wyrazów – istotność rozumianą jako frekwencję.

Wyłonienie słowozbiorów z korpusu to pierwszy etap analizy, najbardziej spektakularny, ale chyba nie najważniejszy. Jak wspomniano wyżej, efektem działania algorytmu LDA jest nie tylko macierz zawierająca proporcje słów w poszczególnych słowozbiorach, ale i druga macierz, na którą składają się proporcje poszczególnych słowozbiorów w tekstach. Dzięki temu można prześledzić obecność danego tematu w korpusie: jego powtarzalność bądź wyjątkowość, jego frekwencję jako funkcję czasu powstania utworu, współwystępowanie tematów w różnych tekstach itd.



Rys. 3. Udział przykładowych słowozbiorów (nr 44, 41, 50 i 45) w omawianych powieściach, ułożonych w kolejności chronologicznej

Na rys. 3a–d przedstawiono obecność kilku przykładowych słowozbiorów w korpusie 100 powieści. Poszczególne teksty zostały ułożone wedle roku ich powstania, dzięki czemu będzie można prześledzić również chronologiczny rozwój niektórych tematów. Na części wykresów (tutaj niepokazanych) ujawniły się słowozbiory bardzo silnie obecne w całym niemal korpusie, gdzie indziej znalazły się takie, które wystąpiły w wielu tekstach, lecz w stosunkowo małym natężeniu, dość częsta była jednak sytuacja, w której jakiś słowozbiór okazał się znakiem rozpoznawczym tylko jednego tekstu i poza nim prawie zupełnie nie wystąpił. Dobrym przykładem takiego endemicznego zestawu wyrazów jest słowozbiór 44 *Stylizacja gwarowa*, reprezentowany przez następujące wyrazy:

kiej, jeno, ino, juści, se, ni, Antek, wsi, la, ledwie, któren, wieś, oczy, chałupy, wszystkie, niby, znówu, Mateusz, kaj, świat, ciągiem, dopiero, Jezus, całą, cicho, zaraz, całkiem, naród, trza, tyła, jał, nieco, dyć, ano, chałupie, abo, czas, choćby, stary, kowal, kobiety, długo, wnet, wolna, świecie, prawie, naraz, ludzie, drugie, izbie, ...

W powyższych wyrazach oczywiście bez trudu rozpoznajemy słownictwo Reymontowych *Chłopów*. Rzeczywiście: jak widać na rys. 3a, słowozbiór ten jest bardzo silnie obecny w jednej jedynej powieści z roku 1908, czyli właśnie w *Chłopach*, i niemal zupełnie nieobecny w pozostałych powieściach. Intuicja podpowiada, że równie endemiczny okaże się słowozbiór 41 *Pieniądz i kapitał*, definiowany przez wyrazy „rubel”, „pieniądze”, „tysięcy”, „interes” (rys. 2d). Spodziewamy się oczywiście, że ów temat będzie obecny niemal wyłącznie w *Lalce* Prusa i w *Ziemi obiecanej* Reymonta, tymczasem widać wyraźnie (rys. 3b), że jest wcale nieźle reprezentowany w kilku powieściach z przełomu XIX i XX wieku oraz w dalszych kilku tekstach publikowanych w latach 30. XX wieku.

Nie bez powodu wszystkie omawiane powieści zostały ułożone chronologicznie na omawianych wykresach. Największe bowiem nadzieje można chyba wiązać z tym, że modelowanie tematyczne okaże się przydatnym narzędziem w językoznawstwie diachronicznym. Użycie algorytmu LDA w badaniach diachronicznych nie jest oczywiście pomysłem nowym: po metodę dość często sięgali historycy nauki, by zbadać, jak tematyka czasopism naukowych zmieniała się przez kolejne dziesięciolecia²⁶, nie stosowano jednak LDA do śledzenia zmian językowych. Trzeba wyraźnie podkreślić, że użyty w niniejszym szkicu zbiór 100 powieści z lat 1850–1940 jest zbyt mały, by wyciągać wiążące wnioski na temat ewolucji języka. Głównym celem jest jednak przedstawienie pewnego sposobu modelowania zmian językowych, który zostanie szerzej omówiony w osobnym szkicu, na podstawie znacznie większego korpusu diachronicznego.

Wyłonione z powieści słowozbiory z rzadka ukazują czytelną tendencję rosnącą bądź opadającą. Większość zresztą jest reprezentowana w niewielkiej liczbie tekstów. Spośród tych, w których trend daje się zaobserwować, najciekawsze wydają się słowozbiory 45 i 50, które można by nazwać: *Czasowniki w czasie przeszłym* oraz *Czasowniki w czasie teraźniejszym* (zob. chmury słów na rys. 2e–f). Jak widać

²⁶ David J. Newman, Sharon Block, *Probabilistic topic decomposition of an eighteenth-century American newspaper*, „Journal of the American Society for Information Science and Technology” 2006, nr 57 (6), s. 753–767; Andrew Goldstone, Ted Underwood, *What can topic models of PMLA teach us about the history of literary scholarship?*, „Journal of Digital Humanities” 2012, nr 2 (1), <http://journal-of-digital-humanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/> [dostęp 15.03.2016].

na rysunkach 3c oraz 3d, pierwszy z nich jest coraz obficiejszy obecny w kolejnych dekadach XX wieku, podczas gdy drugi, zrazu bardzo silnie reprezentowany, zdaje się coraz bardziej zamierać wraz z upływem czasu. W wypadku tych akurat słowozbiorów (czas przeszły *contra* czas teraźniejszy) dużo większe znaczenie ma zmiana stylistyczna – pewnego rodzaju stylistyczny smak epoki wynikający z tematyki powieści – niż rzeczywista zmiana w języku. Wolno jednak zasadnie twierdzić, że w modelowaniu tematycznym kryje się olbrzymi potencjał do śledzenia nie tylko zmian stylistycznych, ale i ewolucji języka.

ZAKOŃCZENIE

W powyższym – z natury swej bardzo pobieżnym – szkicu przedstawione zostały trzy różne metody automatycznego wydobywania słów „istotnych” z korpusu. Za każdym razem procedura opierała się na tym, że ważność danego słowa w jakiejś części korpusu (np. w jednej próbie) była definiowana w opozycji do całego korpusu albo do jego wszystkich pozostałych części (próbek). Wynika z tego, że słowa kluczowe będą się zmieniały w zależności od kontekstu: słowa uznane za ważne dla literatury powieściowej będą inne, gdy kontekstem będzie zbiór prasy, i inne, gdy korpusem referencyjnym będzie np. eseistyka. To samo ograniczenie będzie widoczne w drugiej z przedstawionych metod, tzn. w metodzie Zeta, a także w modelowaniu tematycznym.

Drugie ograniczenie omawianych metod jest takie, że słowa istotne wyłaniane są wyłącznie na podstawie ich frekwencji i współwystępowania, można więc mieć poważne wątpliwości, czy algorytm wybierze „właściwe” słowa. Zdumiewająco jednak przekonujące wyniki omawianych metod – szczególnie zaś modelowania tematycznego – zdają się przeczyć owym obawom. Trafność wyłonionych automatycznie słowozbiorów skłania do jeszcze jednej refleksji: takiej mianowicie, że ukrytą warstwę relacji semantycznych można częściowo zrekonstruować wyłącznie za pomocą frekwencji wyrazów. Ten ostatni wniosek ma niezaprzeczalną wartość w perspektywie różnorodnych dociekań językoznawczych – chciałoby się wierzyć, że niniejszy szkic będzie impulsem do podjęcia takich szerzej zakrojonych badań.

SIGNIFICANT WORDS, KEYWORDS, WORDLISTS – ON STATISTICAL METHODS OF SEARCHING FOR RELEVANT TERMS

Summary

This article discusses automatic extraction of relevant words from sets of texts. The author briefly presents three methods aimed to extract the words from the corpus of words with regard to their frequency, or words whose occurrence next to each other is not random. First, he focuses on the keyword analysis method, then he discusses the Zeta method

developed by John Burrows and Hugh Craig, and the third method covered in the article is the topic modelling method, which is becoming very popular recently, and consists in finding clusters of words co-occurring in similar contexts. Topic modelling was intended for a quick content search in large collections of documents. On the basis of 100 Polish novels, the article presents how this method can be used for linguistic studies.

Trans. Izabela Ślusarek