Słowozbiory „Tekstów Drugich"

Maciej Maryl, Maciej Eder

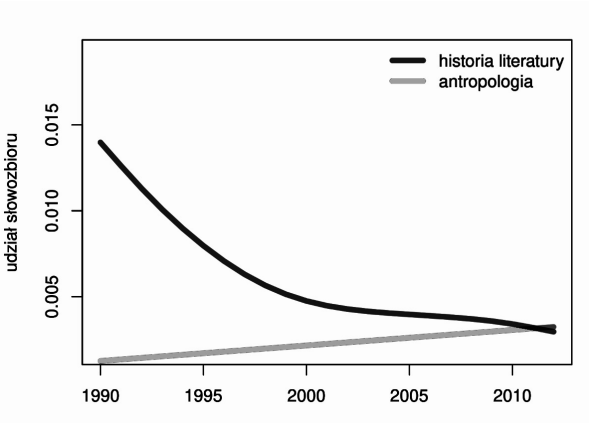
TEKSTY DRUGIE 2023, NR 1, S. 346–364

DOI: 10.18318/td.2023.1.21 | ORCID: Maciej Maryl: 0000-0002-2639-041X
Maciej Eder: 0000-0002-1429-5036

Praca częściowo finansowana z projektu *Cyfrowa infrastruktura badawcza dla humanistyki i nauk o sztuce DARIAH-PL*, Program Operacyjny Inteligentny Rozwój 2014–2020 #POIR.04.02.00-00-D006/20, częściowo zaś z projektu NCN *Wielkoskalowa analiza tekstu i metodologiczne podstawy stylistyki komputerowej*, 2017/26/E/HS2/01019

1. Wprowadzenie

Zacznijmy od wykresu.



Rysunek 1. Częstość słowozbiorów poświęconych historii literatury i antropologii

Wykres obrazuje obecność tematów literaturoznawczych i kulturoznawczych na łamach „Tekstów Drugich" w latach 1990-2012. Nie wdając się w tym miejscu w tłumaczenie metodologii – tym uraczymy czytelników

Maciej Maryl – dr, adiunkt, kierownik Centrum Humanistyki Cyfrowej IBL PAN. Strona: maryl.org. Kontakt: Maciej.Maryl@ibl.waw.pl.

Maciej Eder – dr hab. prof. IJP PAN, Dyrektor Instytututu Języka Polskiego PAN. Strona: maciejeder.org. Kontakt: maciej.eder@ijp.pan.pl.

w dalszych partiach tekstu – spróbujmy przez chwilę rozważyć samo znaczenie wniosków płynących z tego zestawienia. Otóż widzimy tu ilustrację pewnego powolnego procesu, któremu podlega czasopismo – a do pewnego stopnia i cała dyscyplina, polegającego na testowaniu i przysposabianiu nowych metodologii czy dążeniu do zajmowania się szerszymi zjawiskami kulturowymi. Takie tendencje, być może oczywiste dla stałych czytelników, mogą być trudne do dostrzeżenia gołym okiem, zwłaszcza w odniesieniu do mniej wyrazistych tematów i obszarów, o których piszemy w dalszej części tekstu. Proponowana tu metoda pozwala spojrzeć na rzeczywistość kulturową z dystansu jako na zbiór nieciągłych, różnorodnych prądów czy kierunków, które czasem umykają jednostkowej obserwacji.

Dość dobrze tę perspektywę oddaje metafora czytania z dystansu (*distant reading*), zaproponowana przez Franco Morettiego przed dwiema dekadami[1], a niedawno odświeżona przez Teda Underwooda, który ukuł termin „odległy horyzont" (*distant horizon*), służący jednocześnie za tytuł jego ostatniej monografii[2]. Horyzont ów stanowi metaforę przekraczania ograniczeń jednostkowej perspektywy: „Jedna para oczu z poziomu ziemi nie jest w stanie uchwycić krzywizny horyzontu, a tezy ograniczone pamięcią jednostkowego czytelnika nie mogą odsłonić największych reguł organizujących historię literatury"[3]. Tak jak dopiero spojrzenie z wysoka pozwala dostrzec krzywiznę ziemi, tak metody cyfrowe pomagają unaocznić pewne procesy, umożliwiając lekturę w innej skali. Underwood stwierdza, że literaturoznawstwo skupia się na analizie pewnych całostek czasowych – takich jak epoki, prądy, pokolenia czy biografie autorów – jednak ma trudności z analizą procesów wykraczających poza tę skalę. A zatem chodzi tu nie tylko o uwzględnienie bogatszego materiału, lecz także (a może: zwłaszcza) o umożliwienie procesualnego, a nie dyskretnego rozumienia historii literatury, czyli zastąpienie zamkniętych całostek czasowych (nierzadko fastrygowanych ze sobą „przełomami" czy sekwencjami zdarzeń) długim trwaniem[4].

---

1 F. Moretti, *Conjectures on World Literature*, „New Left Review" 2000, no. 1, https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature (24.02.2023).

2 T. Underwood, *Distant Horizons: Digital Evidence and Literary Change*, Chicago University Press, Chicago 2019.

3 Tamże, s. X.

4 Więcej o koncepcji Underwooda w: M. Maryl, *Computational Monograph: Reading and Writing Distant Horizons*, „JLTonline" 2020, vol. 14, no. 2, http://www.jltonline.de/index.php/reviews/article/view/1090/2504 (23.02.2023).

Jeśli spojrzeć ponownie na wykres otwierający ten tekst, zobaczymy ilustrację procesu, jaki zaszedł na łamach „Tekstów Drugich" od początku istnienia pisma, do pewnego stopnia oddającą zapewne ewolucję samego literaturoznawstwa. Chodzi tu oczywiście o stopniowe odchodzenie od problematyki czysto historycznoliterackiej i otwieranie się na inne metody i dyscypliny, zwłaszcza kulturoznawcze. Proponowana w tym tekście metoda pozwala dostrzec te procesy, nawet jeśli wydają się one oczywiste przy analizie dobrze znanego materiału; może jednak posłużyć do automatycznego (a więc: znacznie szybszego niż ręczne wertowanie kolejnych numerów) opisu zawartości czasopisma, z którym nie mieliśmy wcześniej do czynienia. Spojrzenie z dystansu, grupujące teksty według słów kluczowych oddających problematykę lub metodologię, jest spojrzeniem świeżym, od razu sytuującym wnioski w pewnej perspektywie. Słowozbiór[5], czyli grupa słów ze sobą współwystępujących, jest bytem historycznym, uzależnionym od kontekstu i pojawia się z różnym natężeniem w badanym okresie. Takie podejście, ujawniające różne metodologie, pozwala rozpoznać głęboko interdyscyplinarny charakter literaturoznawstwa, które nie stanowi monolitu, lecz grupę zagadnień i podejść czerpiących z innych dyscyplin.

Badania, które tu przedstawiamy, stanowią kontynuację dociekań prezentowanych w pracy *Tekstów świat*, w której Maciej Maryl wykorzystał metadane artykułów ukazujących się w pierwszym ćwierćwieczu istnienia „Tekstów Drugich" (bibliografie i wyniki statystycznych analiz samych tekstów), by opisać trendy w skali makro i przedstawić najważniejsze odniesienia autorów artykułów, budując na ich podstawie model grup intelektualnych w historii pisma[6]. Tekst kończył się zarysem perspektywy analizy tematycznej „Tekstów Drugich", którą podejmujemy w tym miejscu.

Pięć lat, jakie dzielą te dwa teksty, to czas dynamicznego rozwoju cyfrowych badań literackich, zwanych też analityką kulturową. Po epoce zachwytu nieograniczonymi możliwościami nowych metod i uwodzącymi wykresami Morettiego czy Jockersa[7] pojawiły się głosy krytyczne nowego pokolenia

5 Termin „słowozbiór" został zaproponowany w pracy Macieja Edera, *Słowa znaczące, słowa kluczowe, słowozbiory – o statystycznych metodach wyszukiwania wyrazów istotnych*, „Przegląd Humanistyczny" 2016, nr 60 (3). Jest to próba spolszczenia wieloznacznego angielskiego terminu *topic*.

6 M. Maryl, *Tekstów świat. Przyczynek do makroanalitycznej monografii czasopisma literaturoznawczego*, w: *Projekt na daleką metę. Prace ofiarowane Ryszardowi Nyczowi*, Wydawnictwo IBL PAN, Warszawa 2016, https://doi.org/10.5281/zenodo.829923 (23.02.2023).

7 Zob. np. F. Moretti, *Distant Reading*, Verso, London 2013; M.L. Jockers, *Macroanalysis: Digital Methods and Literary History*, University of Illinois Press, Chicago 2013.

badaczy, którzy zakwestionowali podstawy metodologiczne tych badań, przenosząc tym samym na nurt humanistyki zagadnienia związane z kryzysem replikacyjnym nauk ścisłych. Wartość metod statystycznych polegać ma bowiem na pewnej niezależności od subiektywnej obserwacji, o czym świadczy możliwość powtórzenia badania i otrzymanie tych samych wyników przez inny zespół. Replikacja nie jest jednakże możliwa, jeżeli dane oraz kod służący do ich przetwarzania nie zostaną udostępnione, na co w 2017 roku zwróciła uwagę Kathrine Bode w swym głośnym tekście o kryzysie replikacyjnym w humanistyce cyfrowej, krytykując nietransparentne praktyki wspomnianych przed chwilą badaczy[8]. Co więcej, Bode zwraca uwagę, że chodzi tu nie tylko o transparentność, lecz także o wiedzę, czy badany materiał rzeczywiście pozwala na wyciągnięcie danych wniosków, ponieważ – jak powiada badaczka – wytwarzanie przedmiotu dla analiz opartych na danych jest samo w sobie działaniem krytycznym i interpretacyjnym. To w braku odpowiednio przygotowanych materiałów Bode dopatruje się trudności z integracją tradycyjnych metod historycznoliterackich z metodami ilościowymi i postuluje kierowanie się „świadomością bibliograficzną" w doborze źródeł. Kluczem jest zatem starannie dobrany zestaw danych i świadomość ich wad czy ograniczeń.

Od nieco innej strony krytykę poprowadziła Nan Z. Da, która podjęła próbę replikacji wyników kilku prac z literaturoznawstwa cyfrowego. Wnioski ogłoszone w tekście *The Computational Case against Computational Literary Studies* spotkały się z gorącą dyskusją („Critical Inquiry" otworzyło specjalne forum internetowe dla krytycznych komentarzy)[9]. Wśród licznych uwag, często odnoszących się do szczegółów metodologii konkretnych prac, badaczka podniosła nadrzędny zarzut „fundamentalnego niedopasowania wykorzystywanych narzędzi statystycznych do przedmiotu badań"[10]. Według Da na tym właśnie polega kluczowy problem ilościowego literaturoznawstwa: „to, co solidne, jest oczywiste (w sensie empirycznym), a to, co nieoczywiste, solidne nie jest; sytuacja trudna do rozwiązania, biorąc pod uwagę naturę danych

8 K. Bode, *The Equivalence of „Close" and „Distant" Reading; or, Toward a New Object for Data-Rich Literary History*, „Modern Language Quarterly" 2017, no. 78, https://doi.org/10.1215/00267929-3699787 (23.02.2023).

9 N.Z. Da, *The Computational Case against Computational Literary Studies*, „Critical Inquiry" 2019, vol. 45, no. 3, https://doi.org/10.1086/702594 (24.02.2023). Wspomniane forum dostępne tutaj: https://critinq.wordpress.com/2019/03/31/computational-literary-studies-a-critical-inquiry-online-forum/ (23.02.2023).

10 N.Z. Da, *The Computational Case*..., s. 601.

literaturoznawczych i naturę badań statystycznych"[11]. Ta nieadekwatność dotyczy także zagadnienia danych poruszanego przez Bode: „W przypadku literatury szybko ujawniają się problemy niewystarczalności i złożoności danych. Ile jest w ogóle osobnych zbiorów danych literaturoznawczych, które można by ręcznie zanotować, a które byłyby wystarczająco duże, by zastosować metody przetwarzania języka naturalnego?"[12].

W niniejszej pracy staramy się wcielić te postulaty w życie, to znaczy wykorzystać adekwatne metody do dobrze przygotowanego materiału, by zaprezentować solidne wnioski. Redukujemy zatem zasięg pracy do jednego czasopisma literaturoznawczego, by na tym przykładzie prześledzić możliwości i ograniczenia proponowanych metod, ale przede wszystkim po to, by powiązać spojrzenie z dystansu z dokładniejszą interpretacją wyników. Zrywamy tym samym z przyczynkarskim charakterem pierwszej fali cyfrowego literaturoznawstwa, oczekując, iż nasze wnioski okażą się interesujące w oderwaniu od samej metody. Aby ułatwić czytelnikom odbiór, krytyczną analizę i ewentualną replikację uzyskanych wyników, publikujemy dane badawcze i kod wykorzystany do badań[13]. Dlatego też skupiamy się na dobrze przygotowanym korpusie „Tekstów Drugich", pokazując zastosowanie modelowania tematycznego do analizy historii intelektualnej czasopisma literaturoznawczego. Zacznijmy od przybliżenia samej metody i jej aplikacji.

2. Metodologia

W badaniu wykorzystano korpus przygotowany do analiz w przywoływanym wyżej artykule *Tekstów świat*: skany OCR „Tekstów Drugich" z lat 1990-1998 oraz teksty z plików redakcyjnych z lat 1999-2012[14]. Łącznie na korpus składa się 114 numerów od 1/1990 do 6/2012. W chwili gromadzenia materiału ostatnim rocznikiem kompletnym (tj. zawierającym teksty wraz z bibliografią dzieł w nich cytowanych), jakim dysponowali autorzy, był 2012. Korpus z założenia będzie służyć nie tylko do przeszukiwania samych tekstów, ale też do porównania analizy pełnotekstowej z bibliometryczną, dlatego wybrano wyłącznie teksty z bibliografią. Z 1923 tekstów oznaczonych jako posiadających

11 Tamże.

12 Tamże, s. 636.

13 Link do repozytorium: https://github.com/computationalstylistics/slowozbiory_TD/ (23.02.2023).

14 M. Maryl, *Tekstów świat*, s. 446-447.

bibliografię odrzucono 53 – z różnych powodów: brak pliku, prawa autorskie, gatunek nienaukowy (list – 3; ankieta – 8; pożegnanie – 6; archiwalia – 12; rozmowa – 11) czy błędne lub powielone rekordy. Ostatecznie uzyskano 1870 artykułów (7,98 miliona słów), które zlematyzowano do dalszych badań za pomocą Literackiego Eksploratora Maszynowego[15]. Niekiedy w celu uzyskania dokładniejszych wyników teksty dzieli się na mniejsze fragmenty, które są następnie analizowane osobno. Nie zdecydowaliśmy się na tę procedurę, ponieważ zdecydowana większość artykułów publikowanych w „Tekstach Drugich" jest raczej krótka.

Do eksploracji korpusu wybraliśmy stosunkowo nową, choć dobrze już zadomowioną w humanistyce cyfrowej metodę modelowania tematycznego (*topic modelling*) w jej klasycznej odmianie znanej jako Latent Dirichlet Allocation (LDA). Metoda ta, wprowadzona przez Bleia, Nga i Jordana[16], pozwala na znalezienie współwystępujących zbiorów słów (tj. słowozbiorów), które ujawniają (ukryte) relacje semantyczne. „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"[17]. Innymi słowy, algorytm nieznający znaczenia słów zestawia je ze sobą w grupy, w których często współwystępują w danym korpusie.

Metodę modelowania tematycznego opracowano do przeszukiwania i kategoryzowania dużych danych tekstowych, w tym porządkowania dokumentów pod względem podobieństwa tematycznego. Do badania tekstów literaturoznawczych z powodzeniem zastosowali ją Andrew Goldstone i Ted Underwood, którzy analizowali przemiany amerykańskiego literaturoznawstwa na podstawie 21 tysięcy artykułów z „PMLA", „Critical Inquiry", „ELH", „Modern Language Review", „Modern Philology", „New Literary History", „PMLA Review of English Studies" (2014)[18]. W naszej pracy staramy się skupić

15 M. Maryl, M. Piasecki, T. Walkowiak, *Literary Exploration Machine a Web-Based Application for Textual Scholars*, w: *Selected Papers from the CLARIN Annual Conference 2017,* Linköping University Electronic Press, Linköping 2018, http://www.ep.liu.se/ecp/147/011/ecp17147011.pdf (23.02.2023).

16 D.V. Blei, A.Y. Ng, M.I. Jordan, *Latent Dirichlet Allocation*, „Journal of Machine Learning Research" 2003, no. 3 (January).

17 M. Eder (2016), *Słowa znaczące*... Tam też szczegółowe wytłumaczenie metody.

18 A. Goldstone, T. Underwood, *The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us*, „New Literary History" 2014, vol. 45, no. 3, https://doi.org/10.1353/nlh.2014.0025 (23.02.2023).

na przemianach w obrębie jednego czasopisma, testując wykorzystanie metody w języku polskim.

Eksperymenty zostały przeprowadzone przy użyciu dostosowanego skryptu w języku programowania R, uzupełnionego pakietem „stylo"[19] do wstępnego przetwarzania tekstu oraz pakietem „topicmodels"[20] do właściwej analizy. W modelowaniu tematycznym jednym z kluczowych parametrów jest liczba słowozbiorów, które algorytm ma automatycznie wyekstrahować z danych. Na ogół tę liczbę ustawia się arbitralnie, w naszym podejściu wykorzystaliśmy jednakże metodę doboru parametru na drodze eksperymentów, przez wielokrotne trenowanie modelu przy różnej liczbie deklarowanych słowozbiorów[21]. Najlepsze wyniki otrzymywaliśmy dla 120 słowozbiorów[22]; taką też liczbę przyjęliśmy jako parametr przy trenowaniu właściwego modelu.

Drugim parametrem, obok wybranej liczby słowozbiorów, jest podanie pewnej liczby ustalonych *a priori* słów, które zostaną wykluczone z dalszej analizy. Taki zbiór niechcianych słów nazywany jest stoplistą (*stopword list*) i należy do standardowych czynności w komputerowej analizie tekstu. Celem jest usunięcie przyimków, zaimków, partykuł („w", „na" „jego", „się", „czy" itd.) i innych słów wprawdzie fundamentalnych dla składni zdania, ale niemających znaczenia dla treści. Stoplisty układa się zazwyczaj pod kątem danego materiału, by zminimalizować szum pochodzący z danych, taki jak błędy OCR, bardzo częste słowa czy nazwy własne. Na przykład w naszych badaniach usuwaliśmy liczby, by dane wyrażenia nie grupowały się ze sobą tylko dlatego, że pojawiały się na stronie czasopisma o tym samym numerze.

---

19 M. Eder (2016), *Słowa znaczące*...

20 B. Grün, K.Hornik, *Topicmodels: An R Package for Fitting Topic Models*, „Journal of Statistical Software" 2011, vol. 40, no. 13, https://www.jstatsoft.org/article/view/v040i13 (23.02.2023).

21 Zob. T.L. Griffiths, M. Steyvers, *Finding Scientific Topics*, „Proceedings of the National Academy of Sciences" 2004, no 101, suppl_1, http://doi.org/10.1073/pnas.0307752101 (23.02.2023); R. Deveaud, É. SanJuan, P. Bellot, *Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval*, „Document numérique" 2004, vol. 17, no 1, http://doi.org/10.3166/dn.17.1.61-84 (23.02.2023); S. Sbalchiero, M. Eder, *Topic Modeling, Long Texts and the Best Number of Topics: Some Problems and Solutions*, „Quality & Quantity" 2004, vol. 54, no. 4, https://link.springer.com/article/10.1007/s11135-020-00976-w (23.02.2023).

22 Trenowaliśmy niezależne modele dla 30, 40, 50, 60 itd. aż do 300 słowozbiorów i dla każdego z tych modeli obliczaliśmy logarytm z tzw. funkcji wiarygodności (*log-likelihood*). Najwyższą wartość funkcja przyjmowała dla modeli obejmujących 100-150 słowozbiorów, z lekkim wskazaniem na 120 słowozbiorów jako optymalną liczbę. Nasz właściwy model wytrenowaliśmy zatem, przyjmując 120 jako parametr wejściowy.

Przeprowadziliśmy trzy niezależne eksperymenty, w których wygenerowaliśmy modele z trzema rodzajami stoplist i poddaliśmy je weryfikacji (wszystkie stoplisty dostępne są jako dane badawcze do niniejszego tekstu[23]):

A – wariant ze stoplistą ogólną automatycznie pozyskaną przy pomocy powszechnie stosowanej metody TF/IDF, która usuwa słowa występujące powszechnie w danym zbiorze oraz słowa niezwykle rzadkie. Wskazano 45 713 słów, w tym słowa pojawiające się we wszystkich tekstach zbioru, jak słowa funkcyjne (na przykład „się", „lub", „i", „oraz", „ja", „na", „w", „ale") lub użyte mniej niż 5 razy w całym zbiorze (na przykład „pęcina", „pedicure", „patrycjuszowskie", „kukułczy", „niedołęga", „wielbłądzi"). W tej ostatniej kategorii znalazły się również słowa błędnie rozpoznane przez lematyzator (na przykład „wskich", „utożsa"), zawierające błędy literowe, nieliczne zapisy w alfabecie cyrylickim, greckim oraz wszelkie inne zapisy akcydentalne, na tyle rzadkie, że nie wprowadziłyby do modelu nic poza szumem.

B – stoplista z wariantu A uzupełniona nazwiskami z bibliografii dzieł cytowanych w artykułach z korpusu. Dodatkowa lista nazwisk, przejrzana pod kątem duplikatów i literówek, zawierała 2533 pozycje (bez imion).

C – stoplista z wariantu B uzupełniona nazwami własnymi z artykułów wygenerowanymi za pomocą Literackiego Eksploratora Maszynowego[24], który rozpoznał łącznie 18 738 nazw własnych w badanych zbiorze. Do stoplisty zaliczono wszystkie nazwy własne osób (oznaczone przez lematyzator jako *nam_liv*), które występowały w całym korpusie przynajmniej 5 razy – łącznie 5130 nazwisk i imion.

3. Interpretacja

3.1. Wybór wariantu

Słowozbiory z każdego wariantu – A, B i C – zostały poddane procesowi interpretacji i uzgadniania przez obydwu autorów. Do nazwania słowozbioru zazwyczaj wykorzystywano jego najbardziej charakterystyczne słowo (tj. wskazane przez algorytm z największym prawdopodobieństwem), chyba że kontekst wskazywał na takie znaczenie, które łatwiej było oddać innym pojęciem. Ze względu na przyjętą metodę niektóre słowa dobierały się ze sobą w sposób przypadkowy, na przykład dyktowany podobnym zestawem wyrazów używanych w argumentacji. Różnicę między takimi słowozbiorami

23 https://github.com/computationalstylistics/slowozbiory_TD (23.02.2023).

24 M. Maryl, M. Piasecki, T. Walkowiak, *Literary Exploration Machine*...

dobrze obrazuje rysunek 2, na którym z prawej strony mamy słowozbiór C1 „Dekonstrukcja", zaś z lewej artefakt „dyskursywny" C4, tj. zlepek słów często współwystępujących w dyskursie naukowym.

Rysunek 2. Przykłady słowozbiorów

W pierwszej kolejności odrzucono artefakty, to znaczy słowozbiory przypadkowe i trudne do zinterpretowania przez anotatorów. Na przykład czasem grupują się ze sobą wyrazy służące do argumentacji lub wyrażenia w obcych językach. We wszystkich modelach tego typu artefakty stanowiły ok. wszystkich słowozbiorów (odpowiednio 24, 33 i 30 słowozbiorów). Warto podkreślić, że metoda premiuje nietypowe słownictwo, a zatem wyraźnie odznaczają się teksty stosujące specyficzną terminologię i język wyraźnie odrębny od pozostałych, na przykład osobny słowozbiór otrzymały opracowania z zakresu teorii aktora-sieci, neurologii, psychoanalizy i kognitywistyki. W przypadku kierunków mocno odwołujących się do prac obcojęzycznych słowozbiory czasem wiążą się z artefaktami językowymi. Przykładowo, konstruktywizm był często opisywany w źródłach niemieckich, a formalizm – rosyjskich, więc wyraźnie współwystępują z artefaktami niemieckimi w pierwszym przypadku lub tekstami o literaturze rosyjskiej w drugim.

Mimo sporych różnic między wariantami niektóre słowozbiory występowały we wszystkich trzech modelach, jak na przykład „architektura", „judaica", „muzyka", „piekło", „tragedia grecka" czy „uniwersytet". Podstawowa różnica dotyczyła nazw własnych. W pierwszym wariancie nie pomijaliśmy nazwisk (także tych z bibliografii), co sprawiało, że odniesienia do wielkich

nazwisk polskiego literaturoznawstwa (na przykład Janion, Sławiński) czy światowej humanistyki (Foucault, Kristeva) tworzyły wokół siebie odrębne słowozbiory, zagłuszając mniejszych twórców. W wariancie B (bez nazwisk z bibliografii) udało się uniknąć sygnałów nazwiskowych, ale pozostały charakterystyczne imiona, które przejęły funkcję organizującą słowozbiory: Witold (Gombrowicz), Zbigniew (Herbert), Julian (Tuwim), Bolesław (Leśmian), Bruno (Schulz).

Z powyższych względów zdecydowaliśmy się na wariant C (całkowicie pozbawiony nazw własnych), by szukać tematów bardziej ogólnych i wyciszyć sygnały personalne, które mogą zagłuszać jakieś szersze zagadnienie. Jest to świadomy wybór metodologiczny powodowany specyfiką dyskursu naukowego, w którym nazwiska przywołuje się w różnorodnych kontekstach. Eliminujemy je, by skupić się bardziej na znaczeniu.

3.2. Kategorie słowozbiorów

Po wyborze wariantu przystąpiliśmy do kategoryzacji słowozbiorów. Drogą indukcji, tj. oddolnego porządkowania danych na podstawie dostrzeżonych podobieństw, stworzyliśmy grupy słowozbiorów i przypisaliśmy do nich poszczególne elementy.

Tabela 1. Kategorie słowozbiorów

Kategoria słowozbioru	Liczba wystąpień
Temat	51
Artefakt	30
Kierunek badań	15
Metodologia	10
Autorski	5
Rodzaj lub gatunek	5
Epoka lub prąd	4
Razem	**120**

Stosunkowo największą kategorią pozostaje **kategoria tematyczna**, zbierająca wyraziste tematy pojawiające się na łamach „Tekstów Drugich". Słowozbiory pojawiają się na różnym poziomie ogólności, jak choćby zagadnienia bardzo generalne: cywilizacja (C9), świat (C13), erotyka (C25), humor (C92), sfera publiczna (C104), ironia (C77); zagadnienia z pogranicza dyscyplin: ekonomia (C27), prawo (C72), etyka (C74); wydarzenia historyczne lub ograniczone czasowo: Holocaust (C35), powstania (C36), socrealizm (C91), starożytna Grecja (C75), futuryzm włoski (C26), emigracja (C14). Niektóre tematy, zwłaszcza tak stosunkowo wąsko zakrojone jak kamp (C80) czy LGBTQ (C51), czasem trudno odróżnić od kolejnej kategorii, jaką jest kierunek badań.

Do **kierunków badań** zaliczyliśmy słowozbiory odpowiadające różnym prądom, modom i zwrotom w badaniach, widocznym na łamach czasopisma. Warto podkreślić cienką granicę pomiędzy tą kategorią a tematyką (na przykład badania nad Zagładą zakwalifikowaliśmy do tematyki, ale można je pod pewnymi względami potraktować także jako kierunek badań związanych z pamięcią czy problematyką sprawstwa i świadectwa). Wśród kierunków wyróżniamy zatem dobrze ugruntowane metodologie jak: dekonstrukcja (C1) i powiązana z nią krytyka postmodernistyczna (C11), konstruktywizm (C7), komparatystyka (C20), psychoanaliza (C22), fenomenologia (C53) formalizm (C63), hermeneutyka (C94) czy krytyka feministyczna (C120). Znajdziemy tu także mniejsze, ale wyraziste prądy, jak bloomowskie badanie wpływów (C66), teoria aktora-sieci (C90), studia postkolonialne (C106) czy somatyczne (C114).

Do **metodologii** zaliczyliśmy kluczowe działy czy zagadnienia badań literackich i kulturowych, które pod pewnymi względami można potraktować jak kierunki, ale są pojęciami szerszymi. Są to zagadnienia historii literatury (C101), narracja (C52), estetyka (C69), tekst (C95), genologia (C107), wersologia (C110) i interpretacja (C117).

Osobną kategorię stanowi grupa **autorska**. Jak wspominaliśmy wyżej, w wariancie C nie używaliśmy już imion ani nazwisk, co pozwoliło nam patrzeć na w pełni tematyczne ukształtowanie słowozbiorów. Jednakowoż w przypadku kilku twórców nie udało się uniknąć wyraźnych sygnałów, które naprowadzają na konkretną postać, jak Mickiewicz (C18: prelekcja, paryski, mistrz, romantyk, towiańczyk, mesjanizm), Gombrowicz (C19: kosmos, forma, gombrowiczowski, ślub, pornografia), Miłosz (C76: traktat, metafizyczny, dolina, umysł, zniewolić), Schulz (C6: sklep, cynamonowy, materia, tandeta, wiosna, manekin, sanatorium). Tytuły utworów często przywoływane

w pracach stanowią mocny sygnał konsolidujący słowozbiory. Warto podkreślić, że wynika to nie tyle z wyrazistości słów tytułowych, ile częstości ich użycia obok siebie.

Wśród **rodzajów i gatunków** wyraźnie można wyróżnić dramat i szerzej zagadnienia związane z teatrem (C41), lirykę (C41), a także garść gatunków, do którego opisu wykorzystywane jest specyficzne słownictwo: folklor śpiewany (C79), fantastyka (C21) i fantastyka naukowa (C119).

Wyraźnie daje się także wydzielić grupę tematów poświęconych konkretnym **epokom i prądom**: literaturze staropolskiej (C3), pozytywizmowi (C5), romantyzmowi (C98) czy zagadnieniom awangardy (C111).

3.3. Ewolucja słowozbiorów w czasie

Do tej pory omawialiśmy słowozbiory jako statyczne tematy pojawiające się na łamach czasopisma na przestrzeni dwóch dekad. Zupełnie nową perspektywę zyskujemy, spoglądając na żywotność – czy też, jeśli kto woli: frekwencję – poszczególnych słowozbiorów w funkcji czasu. W ten sposób możemy się przyjrzeć ewolucji pisma i tym samym wrócić do wykresu, od którego zaczęliśmy ten tekst.

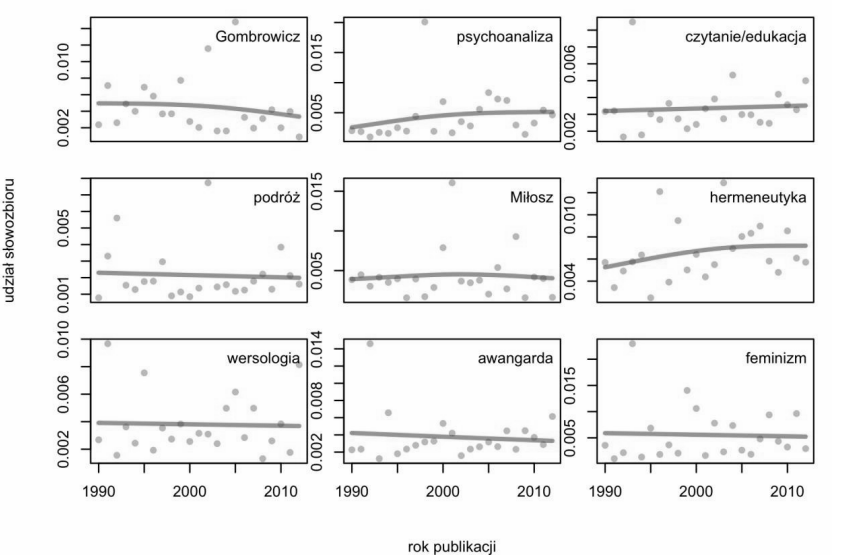
Zanim jednak przejdziemy do wyników, warto poczynić niewielki metodologiczny ekskurs. Otóż końcowym efektem modelowania tematycznego jest, po pierwsze, zestaw słowozbiorów wyekstrahowanych z korpusu tekstów, a po drugie – informacja o udziale poszczególnych słowozbiorów w każdym tekście z korpusu. Pierwszej z tych dwóch informacji mieliśmy już okazję się przyjrzeć: na podstawie prawdopodobieństwa wystąpienia danego słowa w danym słowozbiorze mogliśmy te proporcje zobrazować w postaci graficznej (rysunek 2) i dostrzec w poszczególnych grupach współwystępujących słów takie zagadnienia jak ironia czy studia postkolonialne. Druga część wyniku to informacja o tym, jakie słowozbiory wystąpiły na przykład we wstępniaku numeru 5/2010 (oraz w każdym innym artykule z naszego korpusu). Jeśli zbierzemy tę informację ze wszystkich tekstów opublikowanych w jakimś roku – niech to będzie wspomniany 2010 – to dostaniemy wgląd w to, jakie tematy najchętniej poruszano w omawianym okresie. Stąd już tylko krok, by powyższą operację powtórzyć osobno dla pozostałych roczników pisma, tak żeby cały okres 1990-2012 znalazł się pod analitycznym szkiełkiem i okiem. Dzięki temu możemy dostrzec różne tematyczne mody w ich długim trwaniu, bez ręcznego kartkowania wszystkich roczników pisma.

Należy jednakże pamiętać, że zmiany obecności słowozbiorów w danym roku są relatywne do ich obecności w pozostałych latach. Niektóre słowa są generalnie bardzo rzadkie, a ich ogólna reprezentacja stosunkowo niska (np. słowo „książka" w temacie C63 bibliologia), inne słowa mogą występować nawet wielokrotnie częściej (np. ta sama „książka" jest piętnastokrotnie bardziej obecna w temacie C83 „pisanie", przy czym dla obu tych słowozbiorów „książka" jest wyrazem konstytutywnym – po prostu C83 jest dużo mocniejszy w „Tekstach Drugich" niż C63). Dlatego też różnice diachroniczne między słowozbiorami mogą pojawiać się na różnych skalach. Ewolucję analizujemy na podstawie linii trendu, która wskazuje tendencję, ale bierzemy też pod uwagę same wartości natężenia średniego udziału słowozbioru w danych latach.

Analizując diachronicznie słowozbiory „Tekstów Drugich", możemy określić cztery główne tendencje w ewolucji tematów w czasie: (a) **stałość**, czyli tematy regularnie obecne z podobnym natężeniem, (b) **spadek** – tematy zanikające, (c) **wzrost** – tematy zyskujące na znaczeniu, (d) **sezonowość** – tematy pojawiające się w danym momencie i wyczerpujące się po jakimś czasie. Warto tu oczywiście podkreślić, że różnica między trzema ostatnimi grupami może być traktowana umownie – można założyć, że tematy bardziej znaczące w ostatnich latach reprezentowanych w naszym korpusie mogą w przyszłości zaniknąć i *vice versa*. Koncentrujemy się tu jednak na bieżącym korpusie.

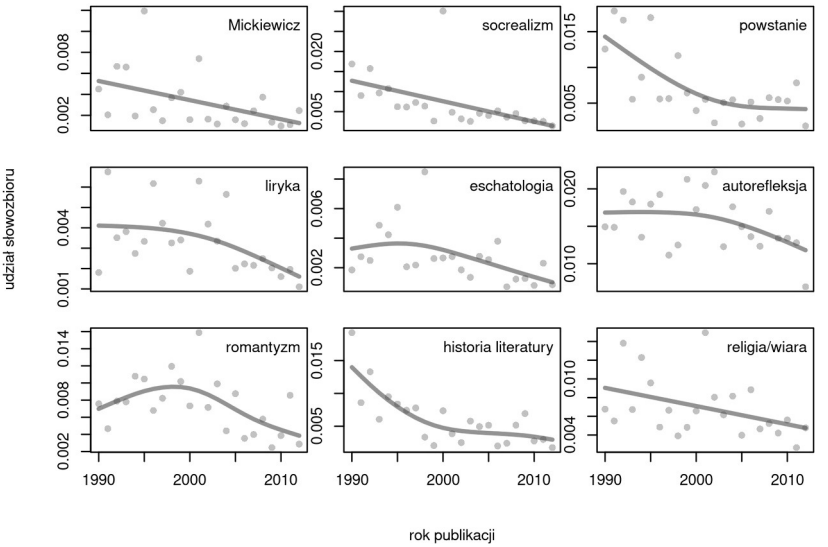
Tematy stałe rzadko charakteryzują się prostą linią trendu idącą przez kolejne lata – jest ona raczej pofalowana, co pokazuje, że zainteresowanie tematem raz się zwiększa, innym razem maleje, ale średnio utrzymuje się na podobnym poziomie przez cały badany okres. Można tu zaliczyć takie kluczowe zagadnienia jak tekst (C95), hermeneutyka (C94), interpretacja (C117), estetyka (C69) czy wersologia (C110), a także kierunki badawcze takie jak formalizm (C63), feminizm (C120), psychoanaliza (C22). Stałe na przestrzeni lat jest także zainteresowanie awangardą (C111), piosenką (C79), czytaniem (C33) czy Gombrowiczem (C19). Wybrane przykłady prezentujemy na rysunku 3.

W niektórych przypadkach linia trendu zdaje się nieznacznie unosić lub opadać, co oznacza powolną zmianę tendencji. Uznajemy jednakże owe zmiany za nieznaczne; mogą zostać zniwelowane lub pogłębione w kolejnych latach. Także obserwacja wyników z kolejnych lat (niebieskie punkty na wykresach) pozwala stwierdzić, czy wzrost lub spadek są efektem incydentalnego zainteresowania, czy trwalszej tendencji. Lepiej to można prześledzić na kolejnym przykładzie tendencji spadkowej.



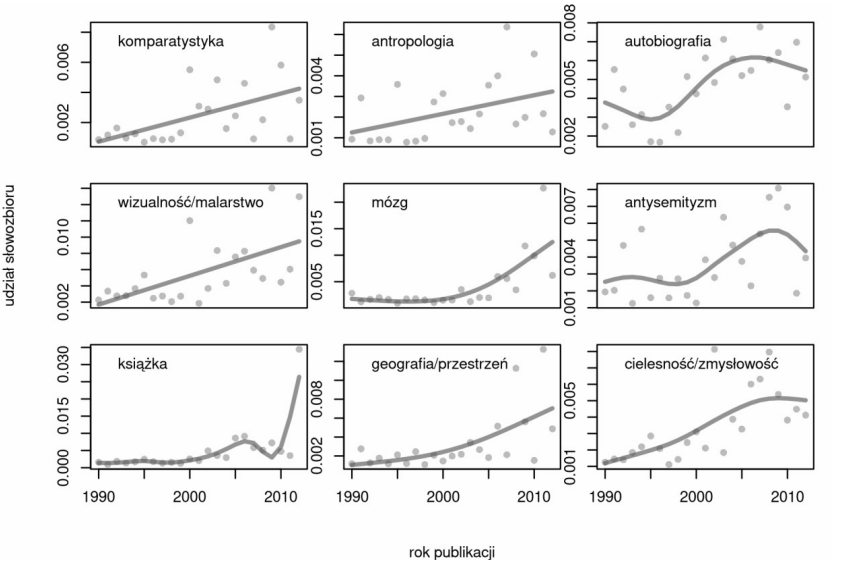
Rysunek 3. Tematy stałe „Tekstów Drugich"

Trend spadkowy dotyczy tematów, które stopniowo przestają cieszyć się zainteresowaniem i znikają z łamów „Tekstów Drugich". Istnieją spadki mocne, rozgrywające się na przestrzeni kilku lat, jak zainteresowanie socrealizmem (C91) czy problematyką powstania i walki (C36), które wyczerpuje się w pierwszej dekadzie istnienia pisma i zapewne może być powiązane z usuwaniem białych plam wiedzy literaturoznawczej w latach dziewięćdziesiątych. Podobny spadek w drugiej dekadzie dotyczy zagadnień związanych z liryką (C42), romantyzmem (98) i religią (C102). Odrębnym, bardzo ciekawym przykładem jest historia literatury (101), która notuje powolny spadek przez kolejne lata istnienia pisma, obrazując głębszy proces zmiany jego problematyki.



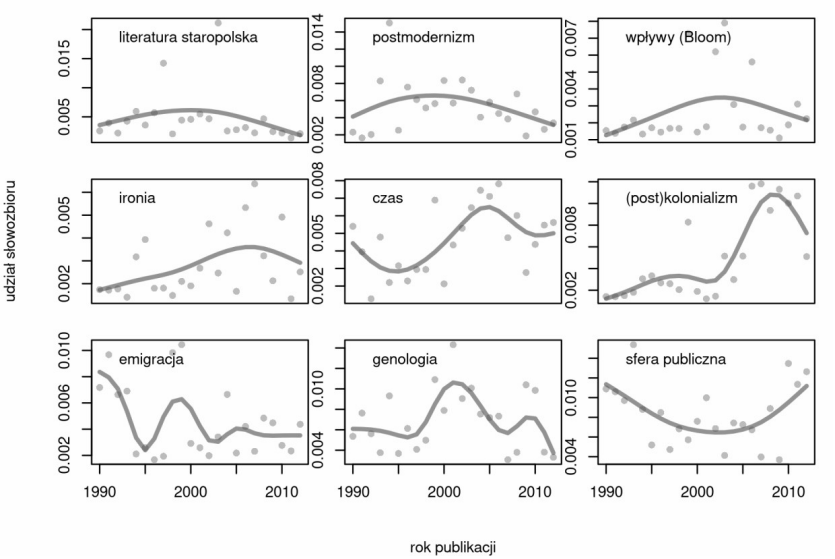
Rysunek 4. Tematy stopniowo wycofujące się z „Tekstów Drugich".

Tendencja **wzrostowa** dotyczy zainteresowania na przykład problematyką dokumentów osobistych (C85), autobiografią (C48), wizualnością (C17, C70), zagadnieniami zagładowymi jak świadek/trauma (C35) czy antysemityzm (C118). Zwraca uwagę zaszczepienie i rozwój na gruncie krajowym takich kierunków jak komparatystyka (C20), krytyka postkolonialna (C106), kognitywistyka (C23) czy teoria aktora-sieci (C90). Interesująco wygląda też temat książki (C44), obejmujący zagadnienia komunikacyjne i medialne – znaczny wzrost pod koniec omawianego okresu pokazuje nagłe zwiększenie zainteresowania tą problematyką. Co ciekawe, tendencja wzrostowa odbija się także w artefaktach – możemy zaobserwować wzrost słownictwa angielskiego (C55) i zwrotów związanych z tłumaczeniami (C113).



Rysunek 5. Tematy zyskujące na popularności w „Tekstach Drugich"

Wreszcie ostatnia tendencja, **sezonowa**, przyjmuje kształt kapelusza – pojawia się i wyczerpuje w czasie omawianych tu dwóch dekad. Taki trend można prześledzić na przykładzie tematu (post)modernistycznego (C11), genologicznego (C107), mocno obecnego w drugiej połowie lat dziewięćdziesiątych i pierwszych latach XXI wieku, czy też tematu ironii (C77) mocno eksploatowanego na łamach pisma w drugiej dekadzie jego istnienia.



Rysunek 6. Tematy sezonowe „Tekstów Drugich".

Są też tematy bardziej sezonowe, które przemijają po kilku latach, jak Bloomowskie wpływy (C66) czy krytyka somatyczna (C114). Warto też odnotować jeden z tematów, który przyjmuje kształt kapelusza odwróconego, czyli słowozbiór poświęcony sferze publicznej, państwu i ideologii (C104). Po okresie istotności w pierwszej połowie lat dziewięćdziesiątych przeżywa renesans od 2010 roku. Warto zauważyć powracającą sezonowość tematu emigracyjnego (C14), który budził duże zainteresowanie w pierwszych i ostatnich latach pierwszej dekady istnienia pisma, by następnie przeistoczyć się w temat stały.

Prezentowane w tej części wykresy i interpretacje są oczywiście tylko pewnym wycinkiem wszystkich uzyskanych przez nas wyników. Te zaś są zaledwie wycinkiem wyników, które uzyskalibyśmy, trenując model tematyczny przy użyciu innych parametrów. Zdajemy sobie sprawę z tego, że czytelnicy otrzymują jedynie wstęp do dalszych interpretacji. Zarazem jednak – i to, sądzimy, jest warte podkreślenia – uzyskane przez nas wyniki są powtarzalne i ukazują się naszym oczom także przy zmienionych parametrach modelu.

4. Konkluzje

Na koniec, bogatsi o wiedzę metodologiczną i rozpoznania dotyczące problematyki badawczej prezentowanej na łamach „Tekstów Drugich", powróćmy do sygnalizowanych na wstępie kwestii wkładu metod cyfrowych w refleksję nad literaturoznawstwem. Wartość tych metod można rozpatrywać przynajmniej na trzech płaszczyznach. Po pierwsze, metody ilościowe wprowadzają szerszą skalę analizy. Intuicje dotyczące problematyki czasopisma konfrontujemy z liczbowymi wskaźnikami opracowanymi na próbie blisko dwóch tysięcy tekstów, które pozwalają nam określić nie tylko podstawowe tematy pisma, lecz także zagadnienia sezonowe, porzucone lub niedawno podjęte. Po drugie, modelowanie tematyczne pozwala nam ująć historię dyscypliny nie tyle jako serię odrębnych etapów, ile ciągły proces, w ramach którego współistnieją różne konfiguracje zagadnień. Po trzecie wreszcie, badanie prądów literaturoznawczych pomaga budować samoświadomość dyscypliny i niejako z góry rozpoznać, które kierunki się w niej zadomowiają, a które ulegają szybkiemu wyczerpaniu.

Wyniki zaprezentowane w tym artykule pokazują tylko jedno z możliwych zastosowań metod cyfrowych do prezentowanego materiału. Obecnie pracujemy nad rozwijaniem tych badań przynajmniej w dwóch kierunkach. Po pierwsze, planujemy powiązać artykuły z bibliografią cytowanych w nich utworów, by przeanalizować sieci relacji i współcytowań, oraz powiązać je z badanymi tu słowozbiorami. Drugi nurt prac, prowadzonych obecnie w ramach projektu DARIAH.lab, ma na celu poszerzenie bazy materiałowej, umożliwiając analizę prądów nie tylko w skali jednego czasopisma, ale też większej grupy tekstów zgromadzonych w Korpusie Dyskursu Literaturoznawczego (1822-2022). Zakładamy, że ta metodologia może być także z powodzeniem stosowana do analizy przemiany dyskursu innych dyscyplin.

Abstract

Maciej Maryl, Maciej Eder
INSTITUTE OF LITERARY RESEARCH, POLISH ACADEMY OF SCIENCES; INSTITUTE OF POLISH LANGUAGE, POLISH ACADEMY OF SCIENCES

Topic Modeling on Second Texts

The article analyzes the thematic structure in the issues of the scholarly journal *Second Texts*, in 1990–2012, using the topic modeling methodology. Based on 1923 articles from this period, the authors obtained and interpreted 120 automatically inferred topics from subsequent issues. The analysis allowed for the identification of permanent and seasonal topics, which appeared with a growing or declining frequency over time. The application of quantitative methods enabled the authors to capture the history of the discipline not so much as a series of distinct phases but rather as a continuous process built from different configurations of co-existing research interests.

Keywords

Teksty Drugie, topic modeling, distant reading, digital humanities, literary studies