Teksty Drugie 2014, 2, s. 90-105

Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii

Maciej Eder

Maciej Eder

Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii

Wstęp

Metody ilościowe nieczęsto znajdują zastosowanie w literaturoznawstwie czy w szeroko pojętej filologii, choć dzięki pracom najpierw formalistów ze szkoły praskiej, a następnie strukturalistów postrzeganie języka, wersyfikacji, stylu itd. jako dającego się modelować systemu znaków stało się istotną perspektywą badawczą. W Polsce przełomowe pod tym względem były prace Jerzego Woronczaka nad wierszem Biernata z Lublina czy nad staroczeską *Kroniką Dalimila*[1], rozprawy Jadwigi Sambor nad bogactwem słownikowym[2], a nade wszystko praca zbiorowa pod redakcją Marii Renaty Mayenowej pod znamiennym tytułem *Poetyka i matematyka*[3].

Początek zastosowania metod ilościowych w badaniach literackich nie przypada jednak na wiek XX. Bardzo charakterystycznym przykładem na humanistyczne

Maciej Eder – adiunkt w Instytucie Filologii Polskiej Uniwersytetu Pedagogicznego w Krakowie oraz w Instytucie Języka Polskiego PAN w Krakowie, zajmuje się literaturą XVI i XVII w., edytorstwem oraz stylometrią. Kontakt: maciejeder@gmail.com

1 J. Woronczak *Z badań nad wierszem Biernata z Lublina*, „Pamiętnik Literacki" 1958 z. 3, s. 97-118; J. Woronczak *Zasada budowy wiersza „Kroniki Dalimila"*, „Pamiętnik Literacki" 1963 z. 2, s. 469-478.

2 J. Sambor *Słowa i liczby*, Ossolineum, Wrocław 1972.

3 *Poetyka i matematyka*, red. M.R. Mayenowa, PIW, Warszawa 1965.

fascynacje językiem i poezją, a zarazem na przekonanie o mierze, liczbie i proporcji jako zasadach rządzących światem, są prace Leona Battisty Albertiego (1404-1472) na temat frekwencji samogłosek w różnych tekstach łacińskich, prowadzące do wniosku, że samogłoski *a*, *e* oraz *y* występują w szczególnym nasyceniu w poezji, pozostałe zaś są wyznacznikiem gatunków oratorskich[4]. Innym ważnym przykładem humanistycznej filologii empirycznej – jeśli można sobie pozwolić na takie określenie – jest rozprawa Lorenza Valli (1407-1457) na temat tzw. *Donacji Konstantyna*, rzekomego dokumentu z IV wieku, w którym cesarz Konstantyn jakoby oddaje Rzym w posiadanie papieży. Valla przeprowadził gruntowną analizę składni, morfologii i semantyki donacji, wykazując niezbicie, że dokument jest fałszerstwem i musiał powstać przynajmniej kilka wieków później[5].

Lorenzo Valla został tu przywołany nie bez powodu. Atrybucja autorska mająca za podstawę statystyczną analizę stylu jest bowiem najbardziej znanym obszarem badawczym łączącym metody ilościowe i filologiczne. Jednym z pionierów tej dyscypliny jest Wincenty Lutosławski, badacz Platona, który wypracował „metodę stylometryczną" (jak sam ją określił) do ustalenia chronologii względnej dialogów platońskich[6]. Od czasów Lutosławskiego metody atrybucji autorskiej rozwinęły się gwałtownie, w ostatnich dekadach coraz bardziej oddalając się od filologii i zbliżając w stronę zaawansowanych technik uczenia maszynowego i sztucznej inteligencji.

Próby komputerowo wspomaganej analizy tekstów *stricte* literackich, mimo że wychodzą z nieco innych założeń teoretycznych, też już podejmowano. Kamieniami milowymi tak rozumianego literaturoznawstwa są rozprawy Johna Burrowsa[7] z jego ideą „computation into criticism" (krytyka literacka bazująca na obliczeniach komputerowych), Franco Morettiego[8] ze słynną koncepcją „distant reading" (analiza literatury bez zaglądania do treści poszczególnych

4 B. Ycart *Alberti's letter counts*, „Literary and Linguisting Computing" (2013), doi: 10.1093/llc/fqt034. First published online: June 22, 2013.

5 H. Love *Attributing authorship: an introduction*, Cambridge University Press, Cambridge 2002, s. 18-19; S.I. Camporeale *Lorenzo Valla i jego traktat o donacji Konstantyna*, przeł. A. Dudzińska-Facca, Wydawnictwo IFiS PAN, Warszawa 1997, s. 72-83.

6 W. Lutosławki *The origin and growth of Plato's logic: with an account of Plato's style and of the chronology of his writings*, Longmans, London 1897; A. Pawłowski, A. Pacewicz *Wincenty Lutosławski (1863-1954). Philosophe, helléniste ou fondateur sous-estimé de la stylométrie?*, „Historiographia Linguistica" 2004 no. 21, s. 423-447.

7 J. Burrows *Computation into criticism: A study of Jane Austen's novels and an experiment in method*, Clarendon Press, Oxford 1987.

8 F. Moretti *Graphs, maps, trees: abstract models for a literary history*, Verso, London–New York 2005; F. Moretti *Distant Reading*, Verso, London–New York 2013.

dzieł), wreszcie praca Matthew Jockersa[9] wprowadzająca pojęcie „macroanalysis" (badania literackie bazujące na analizie bardzo dużej liczby tekstów jednocześnie). Metody stylometryczne były też z powodzeniem stosowane w studiach nad chronologią dzieł literackich[10], w przekładoznawstwie[11], a także do wyłaniania stylistycznych cech gatunków literackich[12].

W niniejszym artykule podejmę próbę adaptacji metod atrybucji autorskiej do badania większych korpusów tekstów literackich. Skoro możliwe jest rozpoznawanie grup utworów napisanych przez jednego autora, to wolno zasadnie sądzić, że w obrębie dużych korpusów będzie możliwe odnajdywanie podobieństw między całymi grupami tekstów czy to zbliżonych pod względem tradycji literackiej, czy to zbliżonych czasem powstania, czy wreszcie wykazujących mniej lub bardziej subtelne zależności intertekstualne. Jak się jednak za chwilę okaże, próby przeskalowania metod atrybucyjnych mogą prowadzić do bardzo niestabilnych (*ipso facto*: niewiarygodnych) wyników. Otóż wbrew pozorom obiektywizm w statystyce nie wynika sam z siebie i wymaga pewnego reżimu metodologicznego przy planowaniu eksperymentu. O tym głównie będzie niniejszy artykuł.

Atrybucja autorska: podstawowe pojęcia

Atrybucja autorska wychodzi z mniej lub bardziej wprost wyrażonego założenia, że istnieje coś takiego jak stylistyczny „odcisk palca" każdego autora, czyli dające się zmierzyć (lub wykryć w inny sposób) jednostkowe cechy języka. Rozszerzając znaną transformacyjno-generatywną teorię lingwistyczną Noama Chomsky'ego[13], można by rzec, że w dowolnym tekście pewne elementy

9 M. Jockers *Macroanalysis: digital methods and literary history*, University of Illinois Press, Urbana–Chicago–Springfield 2013.

10 C. Stamou *Stylochronometry: stylistic development, sequence of composition, and relative dating*, „Literary and Linguistic Computing" 2008 no. 23, s. 181-199; D. Hoover *Modes of composition in Henry James: dictation, style, and 'What Maisie Knew'*, w: *Digital Humanities 2009: Conference Abstracts*, University of Maryland, College Park 2009, s. 145-148.

11 J. Burrows *The Englishing of Juvenal: computational stylistics and translated texts*, „Style" 2002 no. 36, s. 677-699; J. Rybicki *The great mystery of the (almost) invisible translator: stylometry in translation*, w: *Quantitative Methods in Corpus-Based Translation Studies*, ed. M. Oakes, M. Ji, John Benjamins, Amsterdam 2012, s. 231-250; J. Rybicki *Stylometryczna niewidzialność tłumacza*, „Przekładaniec" 2013, s. 61-87.

12 C. Schöch *Fine-tuning stylometric tools: investigating authorship and genre in French classical theater*, w: *Digital Humanities 2013: Book of Abstracts*, University of Nebraska-Lincoln, Lincoln 2013, s. 383-386; M. Kestemont, K. Luyckx, W. Daelemans, T. Crombez *Cross-genre authorship verification using unmasking*, „English Studies" 2012 no. 93, s. 340-356.

13 N. Chomsky *Syntactic structures*, Mouton, The Hague 1975.

należą do systemu językowego i autor nie ma na nie żadnego wpływu (struktura głęboka), inne z kolei są efektem świadomych stylistycznych wyborów autora wyposażonego w kompetencję językową (struktura powierzchniowa). Jeszcze inne elementy ukształtowanej autorsko wypowiedzi – właśnie ów „odcisk palca" – byłyby w tym modelu czymś pośrednim między strukturą głęboką i powierzchniową: nieuświadomionymi nawykami stylistycznymi objawiającymi się np. nadużywaniem pewnych słów czy zwrotów. Te nieuświadomione nawyki stylistyczne można łączyć nie tylko z wykształceniem, oczytaniem, wrażliwością estetyczną czy ogólnie rozumianym bagażem kulturowym poszczególnych autorów, ale też z innymi czynnikami – na przykład stosunkowo prosto daje się zidentyfikować stylistyczne różnice między tekstami pisanymi przez kobiety i przez mężczyzn[14]. Suma wspomnianych powyżej czynników zewnętrznych oraz nawyków pisarskich składałaby się na ów niepowtarzalny „odcisk palca". Z drugiej jednak strony trzeba pamiętać, że milczące założenie o odrębnym „odcisku palca" każdego autora jest nieco naiwne, pisarze bowiem zawsze wyrastają z jakiejś tradycji literackiej, inspirują się nawzajem, wchodzą w relacje intertekstualne ze swoimi poprzednikami – słowem, istnienie indywidualnego stylu w postaci czystej jest raczej teoretyczną mrzonką niż mierzalnym faktem[15]. Taka subtelna sieć powiązań między autorami stanowi problem w atrybucji autorskiej – bo czasem prowadzi do błędnych wyników – ale jednocześnie jest doskonałym punktem wyjścia do wielkoskalowych badań literaturoznawczych. Do tej kwestii, fundamentalnej dla stylometrii jako propozycji badawczej w literaturoznawstwie, przyjdzie mi jeszcze powrócić w dalszej części niniejszego artykułu.

Paradoksem atrybucji autorskiej bazującej na miarach stylometrycznych jest fakt, że taka analiza całkowicie pomija najistotniejsze cechy stylistyczne tekstów literackich: tropy, figury, stylizację, archaizację, ironię *etc.*, skupia się natomiast na takich elementach języka, które z pozoru nie mają nic wspólnego ze stylem. W fundamentalnym studium na temat autorstwa zbioru felietonów pt. *Federalist papers* z 1787 roku Frederick Mosteller i David Wallace wykazali, że znakomitym znacznikiem różnicującym autorów są wyrazy synsemantyczne: rodzajniki, spójniki, przyimki, partykuły i niektóre zaimki osobowe[16]. Szybko zauważono, że te właśnie wyrazy są jednocześnie najczęstszymi leksemami

14 J.W. Pennabaker *The secret life of pronouns: what our words say about us*, Bloomsbury Press, New York etc. 2011, s. 39-60.

15 Zob. empiryczne studium mierzące m.in. wpływ różnego rodzaju relacji intertekstualnych w korpusie: M. Eder *Mind your corpus: systematic errors in authorship attribution*, „Literary and Linguistic Computing" 2013 no. 28, s. 603-614.

16 F. Mosteller, D.L. Wallace *Inference and disputed authorship: the federalist*. Addison-Wesley, Reading (Mass.) 1964.

każdego języka naturalnego, dlatego też współczesna stylometria w zdecydowanej większości polega na analizie występowania najczęstszych słów u badanych autorów. W języku polskim lista wyrazów o największej frekwencji zaczyna się podobnie: *i*, *się*, *w*, *nie*, *na*, *z*, *że*, …

Obraz tego, jak niewiele współczesna stylometria ma wspólnego z tradycyjnie rozumianym stylem, dać mogą powstające w ostatnich latach prace na temat atrybucji publikowane przez badaczy technologii informacji, w których do rozróżniania autorów z powodzeniem stosuje się zbitki dwóch lub więcej wyrazów, relacje składniowe między wyrazami, a nawet najczęstsze zbitki liter (np. dwóch albo trzech)[17]. Tak rozumiane znaczniki „stylu" są zupełnie oderwane od teorii języka: o ich wyborze decyduje wyłącznie kryterium skuteczności w atrybucji. Przykładowe zdanie – niech to będzie początek księgi Genesis z *Biblii królowej Zofii* – zostałoby podzielone na zbitki trzyliterowe w następujący sposób:

> W początce Bog stworzył niebo i ziemię. Ale ziemia była nieużyteczna a prozna, a ćmy były na twarzy przepaści, a duch Boży na świecie nad wodami.
>
> w_p _po ocz czą ątc tce ce_ e_b _bo bog og_ g_s _st stw two wor …

Następnie takie pozbawione znaczenia zbitki liter zostałyby policzone i ich frekwencje porównane z innymi tekstami. Mimo że stosunkowo efektywne[18], takie podejście ma już niewiele wspólnego z rzeczywistymi jednostkami języka naturalnego, dlatego w dalszej części tego studium będę przywoływać wyłącznie przykłady analiz opartych na frekwencjach najczęstszych wyrazów.

Procedura szukania podobieństw między tekstami za pomocą metod wielowymiarowych jest dość dobrze opisana w literaturze przedmiotu[19].

---

17 Zob. np. E. Stamatatos *A survey of modern authorship attribution methods*, „Journal of the American Society of Information Science and Technology" 2009 no. 60, s. 538-556.

18 Studium porównawcze różnych typów znaczników stylu w czterech różnych językach pokazało, że zwykłe frekwencje najczęstszych słów są jednak najskuteczniejszym sposobem rozpoznawania autorstwa, zob. M. Eder *Style-markers in authorship attribution: a cross-language study of the authorial fingerprint*, „Studies in Polish Linguistics" 2011 no. 6, s. 99-114.

19 Zob. np. D. Hoover *Statistical stylistic and authorship attribution: an empirical investigation*, „Literary and Linguistic Computing" 2001 no. 16, s. 421-444; J. Burrows *'Delta': a measure of stylistic difference and a guide to likely authorship*, „Literary and Linguistic Computing" 2002 no. 17, 267--287; P. Juola *Authorship attribution*, „Foundations and Trends in Information Retrieval" 2006 no. 1, s. 233-334; M. Koppel, J. Schler, S. Argamon *Computational methods in authorship attribution*, „Journal of the American Society for Information Science and Technology" 2009 no. 60, s. 9-26.

Na potrzeby niniejszego szkicu wystarczy wspomnieć skrótowo, że różnica w użyciu jakiegoś słowa między dwoma tekstami (np. tekst A częściej używa słowa *się* niż tekst B) zostaje przełożona na sumaryczną miarę odległości. Dzięki akumulacji poszczególnych różnic między kolejnymi słowami (*i*, *się*, *w*, *nie*, ...) sumaryczna odległość między badanymi tekstami też się zwiększa: tym bardziej, im bardziej analizowane teksty różnią się pod względem użycia wyrazów. Procedura obliczania sumarycznej odległości jest przeprowadzana dla każdej możliwej pary tekstów w korpusie. W sytuacji, gdy każdy tekst z korpusu jest konfrontowany ze wszystkimi pozostałymi, za najbardziej podobny zostaje uznany ten, który wykazywał najmniejszą odległość (jest to tzw. metoda najbliższego sąsiada).

Nietrudno zauważyć, że istota atrybucji autorskiej sprowadza się w tym ujęciu do porównania tekstu anonimowego z korpusem referencyjnym i do obliczenia, który z tekstów porównawczych okazał się najbliższym sąsiadem tekstu anonimowego. Jeśli odłożyć na bok szczegóły matematyczne, sama idea leżąca u podstaw stylometrii jest więc zdumiewająco prosta, ale też niepozbawiona pewnej istotnej wady. Otóż najbliższy sąsiad zostanie wskazany bez względu na to, czy korpus porównawczy zawiera próbki tekstowe rzeczywistych kandydatów do autorstwa, czy też składa się z przypadkowo dobranych utworów. Mówiąc obrazowo i z zamierzoną przesadą: jeśli przy dowodzeniu autorstwa *Erotyków* przypisywanych Sępowi Szarzyńskiemu zostanie użyty korpus, powiedzmy, dzieł Mickiewicza, Lechonia i Szymborskiej, to jako najbardziej prawdopodobny autor zostanie wskazany Mickiewicz – bądź co bądź bliższy stylem XVI-wiecznemu autorowi *Erotyków* niż Lechoń albo Szymborska. Ta charakterystyczna przypadłość metod bazujących na mierze odległości będzie miała istotne konsekwencje nie tylko w atrybucji autorskiej, lecz także w próbach zastosowania stylometrii do badania szeroko rozumianych relacji między tekstami literackimi: jakość wyniku eksperymentu będzie bardzo mocno zależna od jakości (oraz zawartości) korpusu porównawczego.

Stylometria w badaniach literatury

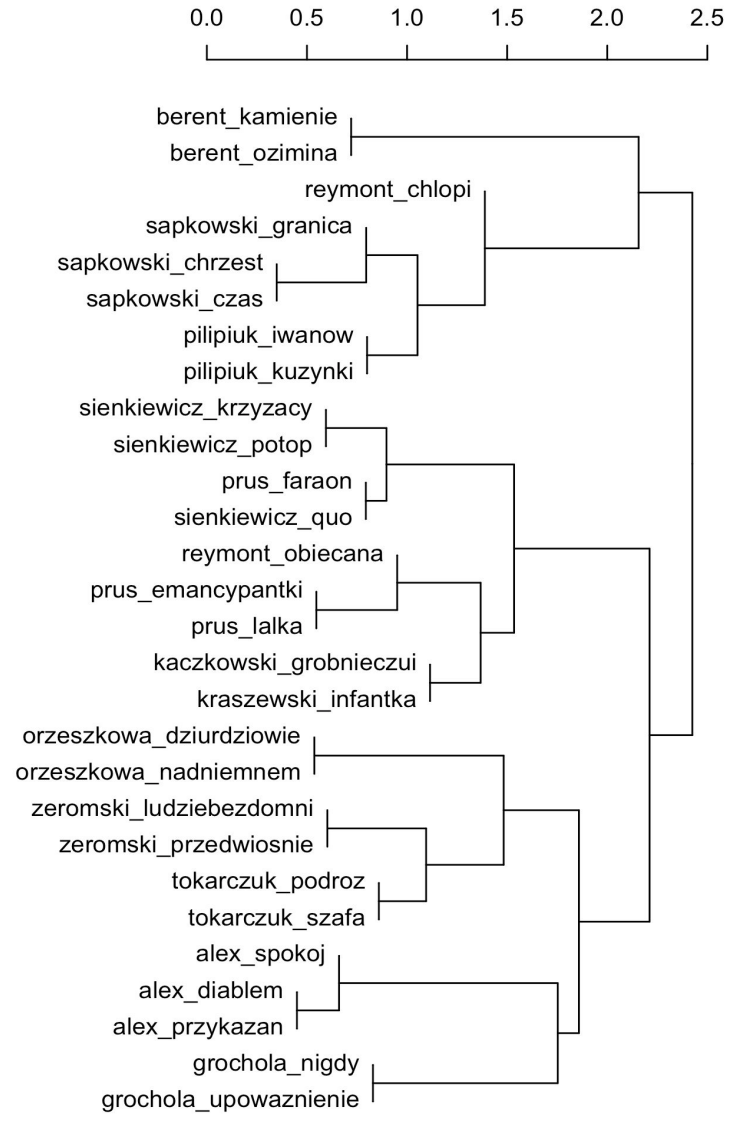
Podstawowe założenie atrybucji autorskiej, czyli zasada najbliższego sąsiada, da się bardzo łatwo uogólnić i przysposobić jako narzędzie w badaniach literackich. Skoro anonimowy tekst poddany analizie okazuje się najbardziej podobny do jakiegoś innego tekstu z korpusu, to tak samo pozostałe teksty z tego samego korpusu są czyimiś najbliższymi sąsiadami, sąsiedzi sąsiadów zaś sąsiadują z jeszcze innymi tekstami, tworząc dość skomplikowaną siatkę wzajemnych powiązań. Zarówno teoria, jak i praktyka pokazują, że tak analizowane teksty na ogół grupują się w większe skupiska: np. autorzy tworzący mniej więcej w tym samym czasie często okazują się wzajemnie podobni,

innym razem daje się zauważyć, że utwory podobne tematycznie wykazują tendencję do grupowania się, klasycznym zaś przykładem przyciągania się podobieństw jest notoryczne grupowanie się powieści sióstr Brontë – Anny, Charlotty i Emily – co stanowi jawny stylometryczny ślad ich wspólnych lektur i wzajemnego czytania/komentowania swoich utworów.

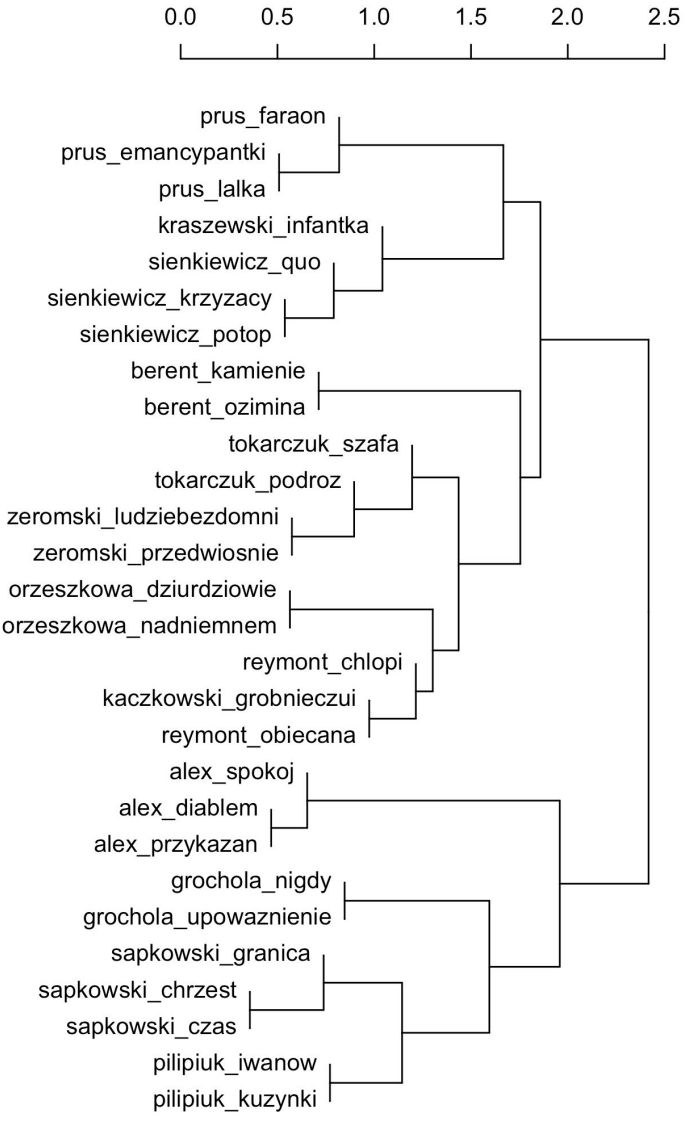
Wolno sądzić, że spojrzenie na historię literatury nie tylko przez pryzmat wnikliwej lektury kilku czy kilkunastu starannie wybranych arcydzieł, lecz także przez analizę stylometryczną wielu tekstów jednocześnie – na przykład kilku tysięcy – może znacząco uzupełnić naszą wiedzę np. o procesie historycznoliterackim. Tak rozumiana stylometria, co warto podkreślić, nie stara się przewartościowywać wiedzy o literaturze czy wprowadzać nowych paradygmatów badawczych. Tak samo jak teleskop jest jedynie narzędziem, użytecznym instrumentem, dzięki któremu widać więcej niż nieuzbrojonym okiem, tak samo statystyka może się stać użytecznym narzędziem, które pozwoli na analizę wielkich korpusów: pytania stawiane literaturze pozostają przecież te same. Istotny problem ze stosowaniem metod ścisłych w humanistyce polega jednak na tym, że na ogół jesteśmy skłonni brać za dobrą monetę wszystko, co obliczy komputer. Obiektywizm wyników bywa jednak nadspodziewanie iluzoryczny.

Na Rys. 1 pokazano przykładowy wynik niewielkiego eksperymentu stylometrycznego: 28 polskich powieści z XIX i XX wieku porównano jedną z metod najbliższego sąsiada, biorąc pod uwagę 100 najczęstszych słów, a następnie przeprowadzono tzw. analizę skupień, która polega na graficznym przedstawieniu podobieństw między poszczególnymi próbkami w postaci dendrogramu (drzewa podobieństw). Interpretacja wykresu polega na odszukaniu „liści" i „gałęzi" oraz ich wzajemnych relacji. Mniejsze „gałęzie" łączą teksty najbardziej do siebie podobne, większe zaś sugerują istnienie bardziej subtelnych więzi stylometrycznych między zgrupowanymi w ten sposób utworami.

Wyniki już na pierwszy rzut oka wydają się ciekawe – np. podobieństwo powieści Joe Alexa do utworów Katarzyny Grocholi musi zastanawiać. Zanim jednak Czytelnik zacznie wyciągać daleko idące wnioski interpretacyjne, warto, by rzucił okiem na Rys. 2, na którym przedstawiono analizę skupień dokładnie tych samych 28 powieści wykonaną przy użyciu dokładnie tej samej procedury i tej samej miary odległości, a jedynym parametrem, który został zmieniony, jest liczba najczęstszych słów wziętych do analizy. Czy Prus jest podobny do Reymonta, czy jednak nie jest? Czy Kaczkowski i Kraszewski zajmują tę samą gałąź wykresu, czy jednak należą do zupełnie innych skupisk? Skoro od metod ścisłych oczekujemy obiektywizmu, to porównanie obu wykresów musi skończyć się niemiłym rozczarowaniem. Gorzej: zarówno 100 najczęstszych



Rys. 1: Analiza skupień 28 powieści polskich, 100 najczęstszych słów.



Rys. 2: Analiza skupień 28 powieści polskich, 300 najczęstszych słów.

słów (Rys. 1), jak i 300 najczęstszych słów (Rys. 2) to wartości dobrane całkowicie arbitralnie. Nie ma żadnego argumentu, który przemawiałby za jedną bądź drugą wartością, nietrudno się też domyślić, że wyniki dla 500 czy 1000 najczęstszych słów przyniosą dalsze przetasowania tekstów przedstawionych na wykresie. Jeszcze gorzej: liczba użytych słów jest tylko jednym z arbitralnie ustalonych parametrów analizy. Zastosowanie innych znaczników stylu (a więc np. wspomnianych powyżej zbitek literowych zamiast najczęstszych słów), innej miary odległości, innego algorytmu budowania dendrogramu i jeszcze kilku pomniejszych parametrów może, choć nie musi, prowadzić do wyłonienia innych połączeń między tekstami. W jaki sposób badacz ma wiedzieć, które parametry dają prawdziwy obraz korpusu, a które prowadzą na stylometryczne manowce? Jak ma wierzyć w obiektywizm wyników, kiedy te przeczą sobie nawzajem?

Nowoczesna statystyka zna różne sposoby radzenia sobie z tym problemem, niektóre z nich są powoli adaptowane na potrzeby stylometrii. Przy analizie skupień – takiej jak demonstrowana powyżej – dość dobrze sprawdza się metoda tzw. drzewek konsensusu wypracowana przez filogenetykę[20], później użyta do badania podobieństw między językami papuaskimi[21], a następnie w stylometrii[22]. W tej metodzie chodzi o automatyczne wygenerowanie dużej liczby tradycyjnych dendrogramów i oszacowanie na ich podstawie dendrogramu uśrednionego: procedura szuka najbardziej stabilnych „gałęzi" drzewa, czyli takich, które najczęściej pojawiają się na poszczególnych dendrogramach, i na tej podstawie rekonstruuje najsilniejsze podobieństwa między tekstami.

Choć problem niestabilności wyników udaje się znacząco ograniczyć przez zastosowanie powyższej metody konsensusowej, nawet tak zaawansowana procedura nie jest w stanie poradzić sobie z innym ograniczeniem, przed jakim staje stylometria wielkoskalowa: gdyby przedstawiony powyżej przykładowy eksperyment został przeprowadzony na kilkuset powieściach, a nie

20 E. Paradis, J. Claude, K. Strimmer *APE: analyses of phylogenetics and evolution in R language*, „Bioinformatics" 2004 no. 20, s. 289-290.

21 M. Dunn, A. Terril, G. Reesink, R.A. Foley, S.C. Levinson *Structural phylogenetics and the reconstruction of ancient language history*, „Science" 2005 no. 309, s. 2072-2075.

22 Metoda drzew konsensusu została przedstawiona w pracy: M. Eder *Computational stylistics and Biblical translation: how reliable can a dendrogram be?*, w: *The translator and the computer*, ed. T. Piotrowski, Ł. Grabowski, Wydawnictwo Wyższej Szkoły Filologicznej, Wrocław 2013, s. 155-170. Zastosowanie metody: J. Rybicki *The great mystery...*; J. Rybicki, M. Heydel *The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish*, „Literary and Linguistic Computing" 2013 no. 28, s. 708-717; K. van Dalen-Oskam *Epistolary voices: the case of Elisabeth Wolff and Agatha Dekken*, w: *Digital Humanities 2013: Book of Abstracts*, University of Nebraska-Lincoln, Lincoln 2013, s. 451-453.

na 28, na wykresie nie udałoby się w sensowny sposób pokazać zależności między tak licznymi próbkami.

Chcąc zobrazować relacje tekstowe w dużych zbiorach danych, trzeba pozbyć się tradycyjnego gorsetu metod wielowymiarowych i sięgnąć po techniki spoza wypróbowanego arsenału stylometrycznego. Wydaje się, że spore nadzieje można wiązać z teorią grafów (teorią sieci). Poniżej zostanie przedstawiona próba zastosowania analizy sieci: celem będzie zarówno uzyskanie stabilności wyników (tj. wyłonienie mocnych, powtarzalnych podobieństw tekstowych), jak i wizualizacja dużych korpusów.

Teoria sieci to gwałtownie rozwijająca się dziedzina wiedzy. Zastosowania analizy sieci znaleźć można niemal wszędzie: i w fizyce jądrowej, i w studiach na temat optymalizacji, i w badaniach mediów społecznościowych. Najsłynniejszy chyba eksperyment opisujący relacje międzyludzkie również opierał się na teorii sieci: dowiedziono, że między dowolną parą osób na świecie jest zaledwie sześć stopni oddalenia, tzn. łańcuch znajomości pośrednich (znajomy znajomego znajomego...) ma tylko sześć ogniw[23]. W badaniach języka i literatury też już próbowano stosować analizę sieci: obrazowano w ten sposób zależności składniowe w języku angielskim[24], czeskim, niemieckim i rumuńskim[25], analizowano współwystępowanie najczęstszych angielskich przymiotników i rzeczowników[26], pokazywano relacje między znaczeniami słów[27]. Do wizualizacji tekstów literackich analiza sieci została użyta przez Jockersa[28]; zaproponowana poniżej[29] metoda stylometryczna jest do pewnego stopnia inspirowana tym ujęciem.

---

23 A.-L. Barabási *Linked: How everything is connected to everything else and what it means for business, science, and everyday life*, Plume, New York 2003; M.E. Newman, A.-L. Barabási, D.J. Watts *The structure and dynamics of networks*, Princeton University Press, Princeton 2006.

24 R. Ferrer i Cancho *The structure of syntactic dependency networks: insights from recent advances in network theory*, w: *Problems of Quantitative Linguistics*, ed. G. Altmann, V. Levickij, V. Perebyinis, RAM-Verlag, Lüdenscheld 2005, s. 60-75.

25 R. Ferrer i Cancho, R.V. Solé, R. Köhler *Patterns in syntactic dependency networks*, „Physical Review E" 2004 no. 69, 051915, s. 1-8.

26 M.E. Newman *Finding community structure in networks using the eigenvectors of matrices*, „Physical Review E" 2006 no. 74, 036104, s. 14.

27 A. Lancichinetti, R. Radicchi, J.J. Ramasco, S. Fortunato *Finding statistically significant communities in networks*, „PLoS ONE" 2011 no. 6, s. 17.

28 M. Jockers *Macroanalysis*..., s. 154-168.

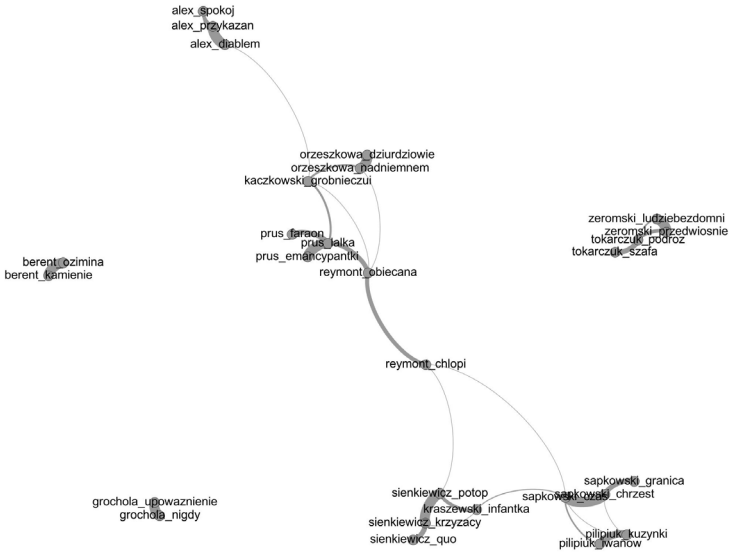
29 W niniejszym artykule przedstawiono zaledwie ogólny zarys metody z pominięciem m.in. założeń matematycznych i sposobu obliczania siły połączeń między tekstami. Szczegółowy opis: M. Eder *Visualization in stylometry: some reliability issues* (w druku).

Teoria sieci (grafów) zakłada, że relacje zachodzące między różnymi zjawiskami można przedstawić jako zbiór połączonych ze sobą punktów-węzłów: jeśli między zjawiskiem A i B zachodzi relacja, to reprezentujące je dwa punkty zostają połączone. Jeśli relacje między badanymi zjawiskami są skomplikowane, to również sieć połączeń między punktami staje się gęsta. Tylko tyle i aż tyle. Cały zaawansowany aparat matematyczny stosowany przy obliczaniu siły powiązań między węzłami to tylko prosta konsekwencja powyższej teorii.

Stosunkowo łatwo można przełożyć relacje między tekstami literackimi na pojęcia teorii grafów. Niech węzłami sieci będą poszczególne teksty z korpusu i niech każdy z tych tekstów zostanie połączony ze swoim najbliższym sąsiadem (tj. tekstem najbardziej podobnym stylistycznie). Tym samym dostaniemy sieć najsilniejszych podobieństw. Niestety już na pierwszy rzut oka widać, że taka sieć przejmie wszystkie ograniczenia dotychczasowych metod, np. analizy skupień. Połączenia między tekstami będą przecież trochę inne dla 100 najczęstszych słów, inne dla 200 słów itd.

Jeżeli jednak dopuścimy myśl, że wyniki dla 100 słów i dla 500 słów pokazują w gruncie rzeczy tę samą rzeczywistość tekstową, ale oglądaną z różnych perspektyw, to możemy podjąć próbę ogarnięcia tej skomplikowanej struktury przez syntezę wielu jednostkowych analiz. Mówiąc obrazowo: jeśli chcemy się czegoś dowiedzieć o katedrze Notre-Dame w Rouen, nie wystarczy obejrzenie jednej fotografii. Kilka różnych ujęć od frontu i od tyłu, w różnych porach dnia, uzupełnione jakimiś sztychami z epoki oraz przede wszystkim słynnymi płótnami Claude Moneta – dopiero wielość perspektyw da nam jakieś wyobrażenie o katedrze. Choć nadal nie jest to wiedza obiektywna, to jednak mając wiele punktów odniesienia, jesteśmy w stanie oddzielić to, co powtarzalne, od tego, co akcydentalne.

Tę samą zasadę możemy zastosować w badaniu tekstów literackich. Jeśli zarysowaną powyżej procedurę znajdowania podobieństw tekstowych powtórzymy dla różnych parametrów eksperymentu, na przykład dla 100, 200, 300, ... 1000 najczęstszych słów, i zsumujemy wszystkie połączenia, które pojawiły się na poszczególnych etapach obliczeń, to efektem będzie sieć rejestrująca wszystkie jednostkowe „fotografie" stylometryczne. Oczywiście pewne połączenia będą się powtarzały, inne okażą się jednorazowe. Im częściej dane połączenie się powtórzy, tym mocniejsza będzie jego siła, co na wykresie zostanie przedstawione grubszą linią. Grubość połączenia jest więc miarą podobieństwa, a pośrednio również miarą stabilności wyników. Rys. 3 przedstawia tego typu sieć; została ona utworzona na podstawie 28 polskich powieści.



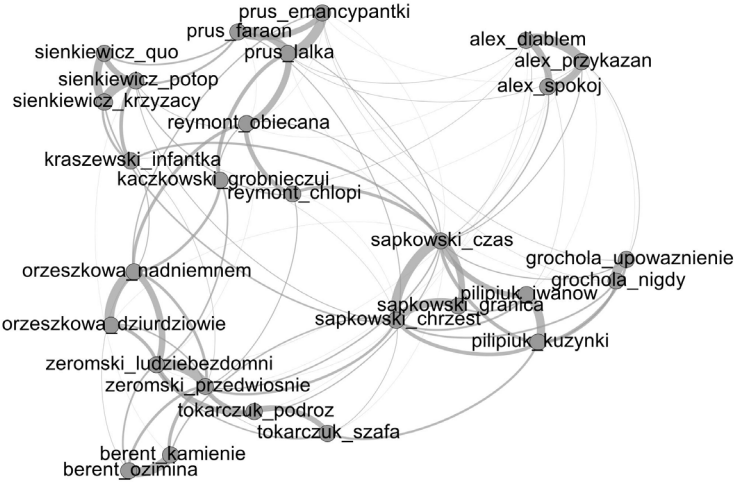
Rys. 3: Sieć stylometryczna 28 powieści polskich, nałożone wyniki dla 100-1000 najczęstszych słów, połączenia między najbliższymi sąsiadami.

Jak widać na wykresie, obie powieści Berenta łączą się wyłącznie ze sobą, tak samo jak powieści Grocholi, Żeromski natomiast okazał się w jakiś stopniu podobny do Tokarczuk. Ciekawy jest również ten rejon sieci, w którym usadowili się z jednej strony Sienkiewicz i Kraszewski, a z drugiej Sapkowski i Pilipiuk (obie podgrupy łączy niezbyt mocne powiązanie Sapkowskiego z Kraszewskim). Trudno się jednak oprzeć wrażeniu, że mimo obiecujących założeń teoretycznych wykres ten niczym szczególnym się nie wyróżnia. Cóż z tego – mógłby ktoś zapytać – że *Ozimina* Berenta okazała się podobna do *Żywych kamieni*? Jeśli badacz literatury miałby odnieść jakąś korzyść z automatycznej analizy korpusu, to pewnie chciałby się raczej dowiedzieć o innych tekstach powiązanych stylistycznie z powieściami Berenta.

W tym momencie dochodzimy do drugiego algorytmu, którego celem jest zobrazowane bardziej subtelnych połączeń tekstowych (pierwszym było sumowanie wielu jednostkowych wyników w jeden obraz). O ile celem atrybucji autorskiej jest odnalezienie wyłącznie najbliższego sąsiada, o tyle w stylometrii wielkoskalowej równie istotne wydaje się pokazanie głębszych relacji sąsiedzkich. Przypomnijmy: w metodach najbliższego sąsiada oblicza się odległość stylometryczną między tekstami – mała odległość oznacza duże podobieństwo i *vice versa* – a następnie na podstawie miar odległości szereguje

się badane utwory od najbardziej do najmniej podobnego. Na przykład dla *Oziminy* Berenta ranking najbardziej podobnych powieści rozpoczynają *Żywe kamienie* (najbliższy sąsiad), potem jest *Przedwiośnie* Żeromskiego, następnie *Ludzie bezdomni*, dalej *Dziurdziowie* Orzeszkowej itd. Najmniej podobne do *Oziminy* okazują się powieści Joe Alexa. Każdy tekst w korpusie ma swój własny ranking sąsiadów, od najbliższego do najdalszego.

Te rankingi sąsiadów mogą zostać wykorzystane przy konstruowaniu sieci. Niech algorytm działa tak, że każdy tekst połączy się z trzema innymi: ze swoim najbliższym sąsiadem, z drugim oraz z trzecim, połączeniami o zróżnicowanej sile (mocnym, średnim i słabym). Efektem będzie gęsta sieć, na której grubymi liniami zostaną oddane silne zależności, słabsze zaś powiązania objawią się jako mniej lub bardziej zwiewne nitki na wykresie – stylometryczne babie lato.



Rys. 4: Sieć stylometryczna 28 powieści polskich, nałożone wyniki dla 100-1000 najczęstszych słów, połączenia między pierwszymi trzema sąsiadami.

Wyniki dla testowego korpusu 28 powieści przedstawiono na Rys. 4. Wykorzystano tutaj oba omówione powyżej algorytmy obliczania siły powiązań[30]. Na wykresie widać dość wyraźnie cztery odrębne grupy tekstów.

30 Wszystkie obliczenia zostały wykonane w środowisku programistycznym R. Implementacja metody przedstawionej w niniejszym artykule jest dostępna w ostatniej wersji pakietu 'stylo'

Pisarzem najbardziej „osobnym" okazuje się Joe Alex, ale nawet on jest połączony cienkimi nićmi podobieństw ze skupiskiem klasyków powieści realistycznej z jednej strony oraz grupą Sapkowski–Grochola–Pilipiuk z drugiej. Interpretację pozostałych połączeń i zgrupowań pozostawiamy Czytelnikowi; godzi się jednak przypomnieć, że korpus zaledwie 28 tekstów nie nadaje się do wyciągania zbyt daleko idących wniosków na temat polskiego powieściopisarstwa. Znacznie bardziej wiarygodny obraz uzyskamy przez porównanie kilkudziesięciu, a jeszcze lepiej: kilkuset powieści. Taką propozycję badawczą przedstawia Jan Rybicki w studium *Pierwszy rzut oka na stylometryczną mapę literatury polskiej*[31].

Zaproponowana powyżej metoda obliczania podobieństw między tekstami literackimi jest próbą zmierzenia się z dwoma problemami jednocześnie: z jednej strony celem było wypracowanie takiego sposobu wizualizacji danych, który umożliwiłby umieszczenie na jednym wykresie bardzo licznych próbek, z drugiej natomiast strony chodziło o przezwyciężenie problemu niestabilnych wyników. Nawet jeśli oba te cele udało się w jakimś stopniu osiągnąć, trudno nie zauważyć, że wyniki analizy nadal będą zależały od kilku arbitralnych decyzji. Czy zatem obiektywne badanie literatury jest w ogóle możliwe?

Obiektywizm jest pewną mrzonką nauk ścisłych, można by rzec Świętym Graalem poszukiwanym od czasów oświecenia, zasadą najpełniej realizowaną przez tzw. metodę naukową (zapewniającą matematyczny formalizm eksperymentu oraz powtarzalność wyników). Nauki humanistyczne cechuje pewnego rodzaju kompleks obiektywizmu, wyrażany i przez Husserla, i przez formalistów szkoły praskiej, wreszcie przez strukturalistów. Obiektywizm w sensie ścisłym, wolno sądzić, pozostanie na zawsze poza zasięgiem badań literaturoznawczych – ze względu na charakter badanego materiału, czyli artystycznie ukształtowanej kreacji literackiej. Celem stylometrii – oczywistym, choć pewnie nigdzie niewyrażonym wprost – jest jednak osiągnięcie obiektywizmu częściowego. Otóż wszystkie testy i procedury statystyczne

(więcej na ten temat zob. https://sites.google.com/site/computationalstylistics/). Do wizualizacji węzłów sieci wykorzystano program Gephi (https://gephi.org). Dość istotną rzeczą w analizie sieci jest dobór algorytmu rozmieszczania węzłów na podstawie ich wzajemnych relacji: niektóre algorytmy próbują wyłonić niewidoczne gołym okiem podgrupy sieci (zgrupowania węzłów), inne rozmieszczają węzły według liczby i siły połączeń wiodących do poszczególnych węzłów. Ten drugi typ układa więc węzły mocno połączone w środku sieci, a na zewnątrz odciąga węzły nietypowe – w sieci stylometrycznej na zewnątrz będziemy na ogół znajdowali teksty awangardowe albo w inny sposób niepasujące do stylu epoki. W centrum znajdziemy teksty najbardziej typowe albo najchętniej naśladowane. W niniejszym artykule stosowano właśnie taki algorytm, zaimplementowany w programie Gephi pod nazwą ForceAtlas2.

31 J. Rybicki *Pierwszy rzut oka na stylometryczną mapę literatury polskiej*, „Teksty Drugie" 2014 nr 2.

przeprowadzane na tekstach są *ex definitione* powtarzalne i weryfikowalne: przy zastosowaniu tych samych korpusów i tego samego zestawu parametrów uzyska się zawsze takie same wyniki. Rzecz jasna, osobną sprawą jest interpretacja tych wyników i tutaj mrzonki o obiektywizmie rozpływają się we mgle mniej lub bardziej błyskotliwych spekulacji.

Abstract

Maciej Eder
PEDAGOGICAL UNIVERSITY OF CRACOW
Scientific methods in literary studies and the traps of apparent objectivism – the case of stylometry

The author discusses several crucial issues related to the use of scientific methods in literary studies, especially the problem of unstable (unreliable) results. With the case study of 28 Polish novels certain limitations of the so-called "cluster analysis" have been presented. In the following part of the article, the author presents the method of calculating textual similarities which is free of such limitations. In order to visualize the results the tools of network analysis have been applied. Even the most advanced techniques of clustering texts are not able to provide absolute objectivism of an experiment.