Teksty Drugie 2014, 2, s. 106-128

Pierwszy rzut oka na stylometryczną mapę literatury polskiej

Jan Rybicki

Jan Rybicki

Pierwszy rzut oka na stylometryczną mapę literatury polskiej

Wstęp

Nie ma lepszego sposobu na przypisanie tekstów ich autorom – oczywiście jeżeli z jakiegoś powodu nie możemy lub nie chcemy korzystać ze stron tytułowych – niż policzenie częstości kilkudziesięciu (lepiej: kilkuset) najczęstszych słów i zastosowanie do niej tej czy innej metody statystycznej (np. analizy skupień), najlepiej bez sprowadzania tych wyrazów do ich form podstawowych (bo w tekście literackim różnica między np. „mówi" a „powiedział" jest stylistycznie bardzo znacząca). Tradycja ilościowych metod stylometrycznej atrybucji autorskiej i stylometrycznej chronologii tekstów sięga jeszcze czasów przedkomputerowych[1], choć oczywiście pojawienie

Jan Rybicki – dr, pracownik IFA UJ, członek Komitetu Wykonawczego European Association for Digital Humanities. Prowadzi badania tekstów literackich stosując metody stylometrii komputerowej. Najnowsze publikacje to *Stylometryczna niewidzialność tłumacza* (2013) i *Collaborative Authorship: Conrad, Ford and Rolling Delta* (2014). Kontakt: jkrybicki@gmail.com

1 Do klasyki ilościowych badań tekstowych należy zaliczyć przede wszystkim Wincentego Lutosławskiego i jego *The origin and growth of Plato's logic: with an account of Plato's style and of the chronology of his writings* (Longmans, Green, London 1897); już wcześniej chronologią Platona zajmowali się w podobny sposób Lewis Campbell (*The Sophistes and Politicus of Plato*, Clarendon Press, London 1867) i Constantin Ritter (*Untersuchunger über Plato. Die echtheit und chronologie der platonischen schriften, nebst anhang: Gedankengang und grundanschauungen von Platos Theätet*, W. Kohlhammer, Stuttgart 1888). Z kolei pomysł na ilościowe ustalenie autentyczności listów św. Pawła miał w 1851 roku sam Augustus de Morgan.

się „mózgów elektronowych", elektronicznych wersji tekstów literackich (i nieliterackich) oraz nowoczesnych metod statystyki komputerowej znacznie ułatwiło takie badania[2] i sprawiło, że podejmuje je coraz więcej badaczy na całym świecie[3].

Nie da się nie zauważyć, że określanie autorstwa na podstawie słów najczęstszych – a więc, co wynika z praw Zipfa[4], nie tylko tych najkrótszych, ale przede wszystkim tych najmniej „znaczących" – przenosi *focus* odczytywania tekstu ze słów takich jak *miłość*, *ojczyzna*, *Polska* czy *duch* albo *dobro*, i przede wszystkim z obszerniejszych związków między słowami (długość i szyk zdania, organizacja tekstu na jeszcze wyższych poziomach hierarchii, relacje intertekstualne), na nieefektowne i częstokroć niezauważalne (a przecież stanowiące często ponad połowę tekstu) słowa „funkcyjne", „synsemantyczne", „gramatyczne", takie jak choćby królujące na szczytach list rangowych *i*, *się*, *w*, *nie*, *na* itd.

Tymczasem okazuje się, że choć analiza stylometryczna bywa często utożsamiana z atrybucją autorską, te same metody często zdają się mówić rzeczy ciekawe (albo przynajmniej tylko: sensowne) o związkach między tekstami

2 Pierwszym zastosowaniem maszyny liczącej (jeszcze nieelektronicznej) w badaniach stylometrycznych było studium Thomasa C. Mendenhalla *A mechanical solution of a literary problem*, „Popular Science Monthly" 1901 nr 60. Przełomowa publikacja to jednak już jak najbardziej komputerowa – i oparta na zliczaniu częstości najczęstszych słów – analiza autorstwa druków ulotnych z początków państwowości amerykańskiej, którą przeprowadzili statystycy Frederick Mosteller i David Wallace, *Inference and disputed authorship: the Federalist*, Addison-Wesley, Reading, 1964. Klasycznym zastosowaniem tych metod w badaniach literackich jest praca Johna Burrowsa *Computation into criticism: a study of Jane Austen's novels and an experiment in method*, Clarendon Press, Oxford 1987.

3 Wspomagane komputerowo badania literaturoznawcze prowadzą m.in. Hugh Craig z Uniwersytetu w Newcastle (Australia): H. Craig, A.F. Kinney *Shakespeare, computers, and the mystery of authorship*, Cambridge University Press, Cambridge 2009; Karina van Dalen-Oskam z Instytutu Huyghensa w Hadze (Holandia): *Names in novels: an experiment in computational stylistics*, „Literary and Linguistic Computing" 2013 nr 28, s. 359-370; David Hoover z New York University (USA): *Corpus stylistics, stylometry, and the styles of Henry James*, „Style" 2007 nr 41(2), s. 174-203; *Stylistics: prospect & retrospect*, Rodopi, Amsterdam 2007; Fotis Jannidis z Uniwersytetu w Würzburgu (Niemcy) i Matthew Jockers z Uniwersytetu w Lincoln (USA): *Macroanalysis. Digital methods and literary history*, University of Illinois Press, Champaign 2013; ten ostatni jest uczniem Franka Morettiego (Stanford University, USA): *Distant reading*, Verso, New York 2013; *Graphs, maps, trees: abstract models for a literary history*, Verso, New York 2005.

4 Najbardziej znanemu prawu Zipfa („frekwencja danego słowa w danym korpusie znaków języka naturalnego jest odwrotnie proporcjonalna do jego miejsca w liście rangowej słownictwa tego tekstu") towarzyszą dwa inne, równie ważne z perspektywy badań takich jak te prezentowane w niniejszym artykule: „im częstsze słowo, tym mniej konkretne znaczenie" i „im krótsze słowo, tym częściej występuje" (por. G.K. Zipf *Human Behaviour and the Principles of Least Effort,* Addison-Wesley, Boston 1949).

różnych autorów. Jeżeli zastosowana metoda jest skuteczna, pojedyncze teksty autora A grupują się razem; osobną grupę stanowią teksty autora B… Jednak kiedy do badanego korpusu dodamy jeszcze równie pięknie łączące się teksty autora C, będą one zwykle bliższe (na wykresie) tekstom – na przykład – A. Czy to oznacza, że autorzy A i C piszą „podobniej" do siebie niż A i B? Czy też jest to zwykły artefakt metody?

Najlepiej zacząć od eksperymentu. Poniżej prezentuję wyniki analizy wzajemnych podobieństw stylometrycznych między tekstami polskich pisarek i pisarzy, bazującej na częstościach najczęstszych słów.

Materiał

Badania przeprowadziłem na korpusie około 500 polskich tekstów literackich ze znaczną przewagą wszelkich podgatunków powieściowych z lat od 1775 (*Mikołaja Doświadczyńskiego przypadki*) do chwili bieżącej (koniec 2013 roku). W korpusie znaleźli się ponadto nieliczni reprezentanci poezji epickiej (*Pan Tadeusz*, *Beniowski* i *Zamek kaniowski*) oraz innych gatunków prozatorskich – przede wszystkim zbiory opowiadań (Schulz, Odojewski); te zostały włączone do korpusu z powodu niedostępności (lub wręcz nieistnienia) dłuższych tekstów literackich autorów ważnych z punktu widzenia historii literatury polskiej. Bazując na pracy Edera[5], w celu zminimalizowania wpływu rozmiarów utworów na wyniki analiz unikałem tekstów poniżej 10 000 słów; najkrótszym tekstem w korpusie jest więc powieść Anny Nakwaskiej, *Aniela, czyli ślubna obrączka* z roku 1831 (13 571 słów), najdłuższym – *Lód* Jacka Dukaja z roku 2007 (391 104).

Wiek XVIII reprezentowany jest – niestety! – jedynie przez wspomniane już dzieło Ignacego Krasickiego; wiek XIX – już przez 111: od *Malwiny* po *Krzyżaków*, *Argonautów* i *Komorników*. Z wieku XX pochodzi 310 tekstów od *Popiołów*, *Próchna* i *Na srebrnym globie* po *Panią jeziora* czy *Dziewięć*. Lista 82 utworów z wieku XXI rozpoczyna się od *Pod mocnym aniołem* i kończy m.in. *Ostatnim rozdaniem*. Przy sporządzeniu korpusu korzystałem przede wszystkim z darmowych źródeł internetowych i z płatnych e-booków, konwertowanych następnie do zwykłego formatu tekstowego; tylko nieliczne książki zostały zeskanowane i poddane optycznemu rozpoznaniu znaków. To pozwoliło na stosunkowo szybkie sporządzenie korpusu, ale równocześnie wykluczyło możliwość użycia wielu tekstów, które zapewne powinny znaleźć się w takiej „reprezentatywnej" czy *horribile dictu*, „kanonicznej" próbce literatury polskiej.

5 M. Eder *Does size matter? Authorship attribution, small samples, big problem*, „Literary and Linguistic Computing", publikacja online, 14.11.2013. http://llc.oxfordjournals.org/content/early/2013/11/14/llc.fqt066.full?sid=5d729c48-0b31-480f-8d69-focde7e195de (dostęp: 1.01.2014).

Jeżeli więc ośmielam się nazywać ten wybór reprezentatywnym, to z zastrzeżeniem, że jest to reprezentatywna próbka materiałów dostępnych obecnie w wersji elektronicznej – ale równocześnie skorzystałem z tego pretekstu, by włączyć do korpusu pozycje z repertuaru popularnego (by nie rzec – brukowego) i/lub młodzieżowego. Natomiast w sposób bardzo nieproporcjonalny potraktowałem literaturę *science fiction* i *fantasy*, która bezapelacyjnie króluje w polskich repozytoriach internetowych; z powodów chyba nie do końca zbadanych to ulubione podgatunki osób nałogowo/zawodowo związanych z komputerami – tych wszystkich informatyków, matematyków i fizyków cząstek elementarnych.

Oczywiście teksty zostały poddane obróbce korektorskiej; natomiast zgodnie z założeniami Craiga i Whippa[6] nie zastosowałem ujednolicenia pisowni tekstów pochodzących z różnych epok; ten problem został zminimalizowany przez zastosowanie wysokich wartości *cullingu* (o czym poniżej).

Metoda

Analizę ilościową przeprowadziłem na ciągach częstości najczęściej występujących słów – w tym przypadku tych, które znalazły się na pierwszych 100-1000 pozycjach list rangowych. W celu wyeliminowania bezpośredniego wpływu tematyki utworów na wyniki analizy zastosowałem *culling* na poziomie 90 i 100% – oznacza to, że w analizie wykorzystano tylko te słowa, które wystąpiły równocześnie albo w 90% wszystkich tekstów, albo we wszystkich tekstach korpusu. W praktyce okazało się, że wszystkie teksty miały nie więcej niż 107 wspólnych słów i tylko 827 słów, które występowały w 9 na 10 tekstów – i właśnie takie były graniczne parametry przedstawionej analizy. Obie listy słów umieściłem w Dodatkach do niniejszego tekstu; warto zauważyć, że przeważają tam wyrazy odpowiedzialne za sposób prowadzenia narracji i kształt dialogów; brak natomiast słów bezpośrednio związanych z treścią utworów – nie wspominając już o imionach i nazwiskach bohaterów czy o nazwach własnych (choćby polskich miast). Co najważniejsze, takie ograniczenie – całkowicie niearbitralne, bo regulowane wyłącznie i automatycznie przez statystykę – likwiduje (jak już wspomniano powyżej) prosty wpływ różnic w pisowni „między dawnymi a młodszymi laty" na wyniki eksperymentu[7].

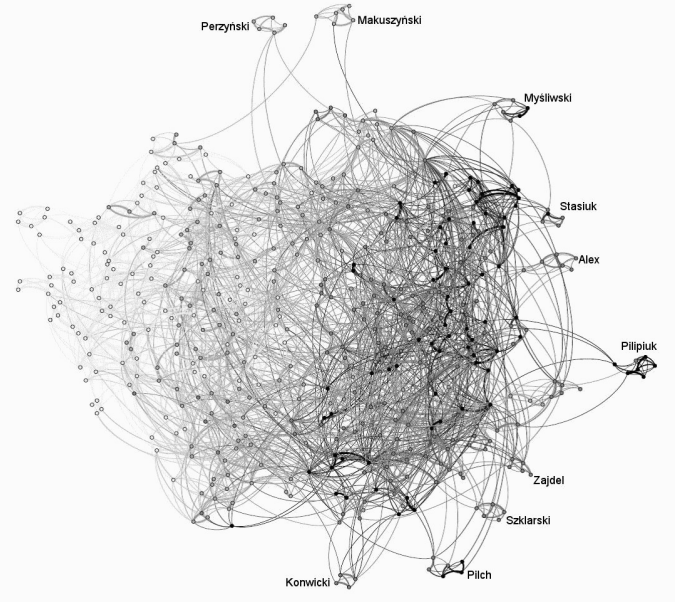
6 H. Craig, R. Whipp *Old spellings, new methods: automated procedures for indeterminate linguistic data*, „Literary and Linguistic Computing" 2010 nr 25(1), s. 37-52.

7 Więcej szczegółów na temat samej metody i jej teoretyczno-statystycznych podstaw przedstawia zamieszczony w tym samym tomie tekst Macieja Edera *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii*.

Wszystkie te procedury – od wczytania tekstów elektronicznych po analizę statystyczną – zostały przeprowadzone za pomocą jednego pakietu „stylo" (wersja 0.5.3) stworzonego przez Macieja Edera przy niewielkim współudziale dwóch pozostałych autorów[8]; pakiet ten jest przeznaczony do środowiska programowania statystycznego R[9]. Natomiast do wizualizacji sieciowej wyników uzyskanych przez „stylo" posłużył open-source'owy program GEPHI[10].

Wyniki

Już pierwszy rzut oka na wyniki analizy sieciowej słownictwa tekstów zawartych w korpusie pozwala zauważyć dość wyraźny sygnał chronologiczny (Wykres 1). Po lewej stronie wykresu przeważają jasnoszare punkty reprezentujące



Wykres 1. Sieć podobieństwa stylometrycznego 503 powieści polskich; sygnał chronologiczny i zjawiska marginalne.

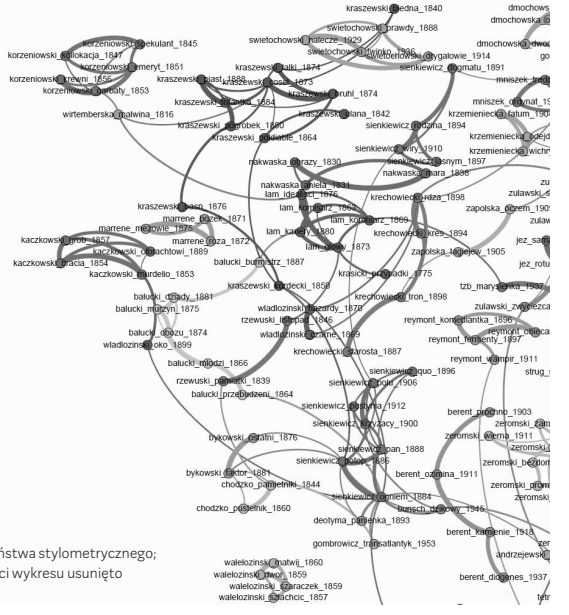
8 M. Eder, M. Kestemont, J. Rybicki *Stylometry with R: a suite of tools*, w: *Digital Humanities 2013: Conference Abstracts*, University of Nebraska-Lincoln, Lincoln 2013, s. 487-489.

9 R Core Team *A language and environment for statistical computing*. R Foundation for Statistical Computing, Wien 2013, http://www.R-project.org/ (dostęp: 1.1.2014).

10 M. Bastian, S. Heymann, M. Jacomy *Gephi: an open source software for exploring and manipulating networks*. International AAAI Conference on Weblogs and Social Media 2009.

literaturę polską do końca XIX wieku. Czarne punkty to dzieła wydane po roku 2000; między nimi rozciąga się szara strefa literatury XX-wiecznej; w niej jednak widać z kolei wyraźny podział na pierwszą (nieco jaśniejszą) i drugą (nieco ciemniejszą) połowę wieku XX. Oczywiście strefy sąsiadujące ze sobą chronologicznie dość silnie na siebie zachodzą, ale po pierwsze, zjawisko to nie dotyczy literatury najwcześniejszej i najnowszej; po drugie, ewolucja od tekstów wczesnych do późnych jest wyraźna i zorientowana wzdłuż poziomej osi wykresu.

O ile makrostruktura tej swoistej literackiej mapy jest zdeterminowana chronologicznie, o tyle jej mikrostruktura ujawnia siłę sygnału autorskiego. Na przykładowym wycinku sieci (Wykres 2) najsilniejsze połączenia (najgrubsze linie) występują między tekstami tych samych autorów, często przy zachowaniu zróżnicowania siły tych połączeń. I tak w twórczości Kaczkowskiego największe podobieństwo stylometryczne występuje między *Murdelionem* i *Braćmi ślubnymi* – może dlatego, że ukazały się w odstępie zapewne roku, a mimo że tylko *Murdelio* należy do cyklu *Ostatni z Nieczujów*. Z kolei dość wczesna, bo należąca jeszcze do wołyńskiego okresu twórczości swego autora powieść obyczajowa *Całe życie biedna* nieco odbiega od późniejszej, przeważnie historycznej twórczości Kraszewskiego. Trudno jednak mówić o podziałach gatunkowych czy podgatunkowych, skoro równie daleko jest do *Króla Piasta* czy *Hrabiny Cosel* równie historycznym *Starej Baśni* i *Kordeckiemu*.



Wykres 2. Wycinek sieci podobieństwa stylometrycznego; sygnał autorski. Dla przejrzystości wykresu usunięto najsłabsze połączenia.

Inni autorzy w przedstawionym wycinku nie pozostają już jednak w takiej samej *splendid isolation*: *Malwinę* Wirtemberskiej łączy stylometryczne podobieństwo o tej samej sile z *Garbatym* Józefa Korzeniowskiego, *Hrabiną Cosel* Kraszewskiego i *Idealistami* Jana Lama. Jeszcze silniej przyciągają się *Murdelio* z *Okiem proroka* Władysława Łozińskiego – do tego stopnia, że opisy perypetii Hanusza Bystrego mają stylometrycznie znacznie mniej wspólnego z innymi dziełami młodszego z braci-pisarzy; tym różnią się więc dwaj Łozińscy od sióstr Brontë, których podobieństwo stylometryczne jest znaczne i zachowuje się nawet w przekładach na inne języki[11]. Zresztą ofiara pojedynku z Karolem Cieszewskim i jego *Zaklęty dwór* czy *Czarny Matwij* też izolują się od stylometrii innych autorów.

Równocześnie jednak sygnał chronologiczny nie jest szczególnie zauważalny w ramach twórczości pojedynczych autorów. Prostą ewolucję cech stylometrycznych u jednego pisarza – taką, w której teksty układałyby się w łańcuch chronologiczny – zaburza fakt, że w większości przypadków utwory jednego autorstwa powiązane są między sobą wielokrotnie. Jednym z nielicznych wyjątków od tej reguły jest stylometria Wacława Berenta, którego *Próchno* (1903) rozpoczyna ewolucyjny szereg przebiegający przez *Oziminę* (1911) i *Żywe kamienie* (1918), a kończący się *Diogenesem w kontuszu* (1937). Być może to nie przypadek, bo właśnie o Berencie pisze Miłosz, że „zdołał w paru powieściach zawrzeć kolejne etapy swojego zmieniającego się poglądu na świat"[12]. Skoro ewolucja poglądów, to może i cech stylometrycznych?

Zupełnie inaczej przedstawia się stylometryczny obraz twórczości Sienkiewicza – ta bowiem rozpada się na dwie niemal odrębne części. W górnej części omawianego wycinka wykresu znalazły się obyczajowe (i zwykle niżej oceniane, i na ogół późniejsze) utwory pierwszego polskiego noblisty; na dole za to gromadzą się największe osiągnięcia amatora licznych Marii – jego powieści historyczno-przygodowe, i to i te pierwsze, i ostatnie (łącznie z mniej udanym *Na polu chwały*). Owe dwa osobne światy łączy tylko jedna cienka linia prowadząca właśnie od najsłabiej przyjętej (i przez samego autora niepoważanej) opowieści sprzed wiktorii wiedeńskiej, *Na polu chwały*, do kompletnego fiaska pisarza, *Wirów*. Trudno powiedzieć, czy stylometria odzwierciedla w ten sposób sąsiedztwo chronologiczne obu powieści (odpowiednio, 1910 i 1906), czy właśnie ich niski poziom literacki, czy wreszcie podobieństwo tematyczne – w końcu z „górnych," obyczajowych powieści właśnie *Wiry* są równocześnie najbardziej polityczne. Oczywiście podobieństwo stylometryczne sienkiewiczowskich romansów przygodowych byłoby najłatwiej przypisać daleko idącej

11 J. Rybicki *Stylometryczna niewidzialność tłumacza*, „Przekładaniec" 2013 nr 27, s. 61-87.

12 Cz. Miłosz *Historia literatury polskiej do roku 1939*, Znak, Kraków 1998, s. 427.

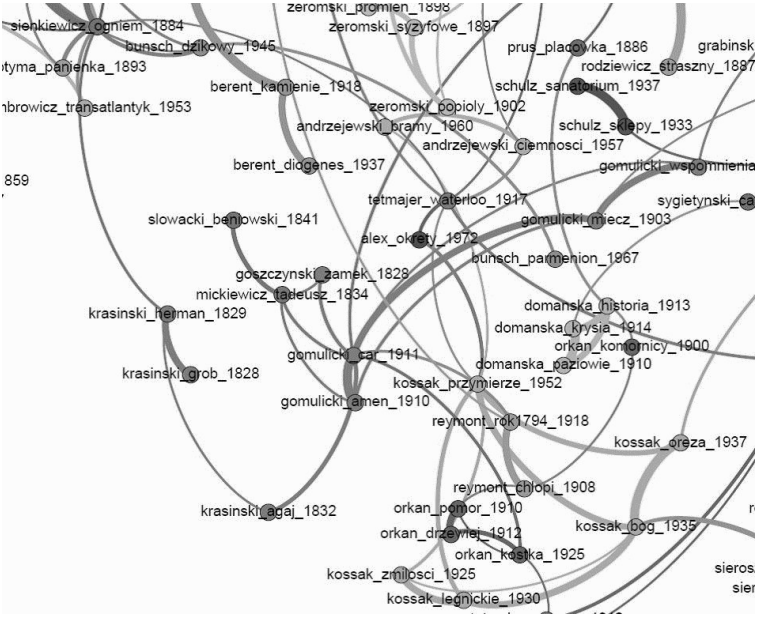
stylizacji Trylogii, *Krzyżaków* i *Quo vadis*, gdyby nie dwie sprawy. Po pierwsze, *Ogniem i mieczem*, *Potop* i *Pan Wołodyjowski* to stylizacja na staropolszczyznę XVII-wieczną; *Krzyżacy* to góralszczyzna udająca średniowiecze; *Quo vadis* przemawia językiem polskich przekładów literatury łacińskiej – więc nie jedna stylizacja, a trzy. Po drugie, trudno mówić o stylizacji języka *W pustyni i w puszczy*, nawet jeżeli jest to polska powieść stylizowana na angielską literaturę kolonialną – i nawet jeżeli „prawdziwym" językiem tej powieści jest… angielski[13].

Ciekawie przedstawiają się związki dwóch sienkiewiczowskich skupisk z dziełami innych pisarzy. *Bez dogmatu* łączy się (poza opisywanym wycinkiem) z debiutem powieściowym Nałkowskiej (*Kobiety*, 1906), która potem oddala się znacznie od sienkiewiczowskich wzorów. Więcej interesujących rzeczy dzieje się jednak wokół „dolnego" Sienkiewicza. Silne podobieństwo stylometryczne między *Potopem* i *Kordeckim* musi stać się okazją do przypomnienia, że analiza została przeprowadzona wyłącznie na słowach najczęstszych (i w dodatku występujących we wszystkich lub prawie wszystkich tekstach w korpusie) – nie zaś na wyrazach, które miałyby jakikolwiek związek znaczeniowy z dwiema wizjami obrony Jasnej Góry przed Szwedami. *Kordecki* jest zresztą też podobny (tylko trochę mniej) do *Ogniem i mieczem* – obie te części Trylogii (wraz z *Krzyżakami*) przeciągają na stronę wieku XIX jak najbardziej XX-wieczny (ale przecież też historyczny i też stylizowany językowo) *Dzikowy skarb* Bunscha; bardzo silnie grawitują do siebie *Ogniem i mieczem* z *Panienką z okienka* Deotymy. Ze stylometrycznego punktu widzenia można upatrywać genezy języka Trylogii nie tylko (o czym już była mowa) w twórczości Kraszewskiego, lecz również w romansach historycznych Piotra Jaksy Bykowskiego; nie da się też przejść do porządku dziennego nad linią łączącą *Ogniem i mieczem* z *Władysławem Hermanem* Zygmunta Krasińskiego. Ten obraz interakcji Trylogii z szerszym repertuarem polskiej powieści historycznej dopełnia (poza wycinkiem) podobieństwo do *Waterloo* Przerwy-Tetmajera (1917).

Skoro wspomnieliśmy już o jednym z Trzech Wieszczów… Wielka to szkoda, że tylko Krasiński popełniał powieści; z pewnym niepokojem o stabilność rodzajową korpusu włączyłem doń jeszcze *Pana Tadeusza* (którego najlepiej znany angielski przekład – pióra wybitnego amerykańskiego slawisty George'a R. Noyesa – pisany jest prozą i czyta się świetnie, zupełnie jak powieść) i *Beniowskiego* – *Maria* okazała się zbyt krótka; kryterium obszerności tekstu spełnił za to *Zamek kaniowski* (ponad 18 tysięcy słów). Polscy romantycy znaleźli się w tej samej okolicy naszej sieci (Wykres 3), ale bezpośredni związek łączy tylko Mickiewicza, Słowackiego i Goszczyńskiego – za mało danych,

13 J. Rybicki *Angielskie przekłady „W pustyni i w puszczy"*, w: *Wokół „W pustyni i w puszczy". W stulecie pierwodruku powieści*, red. J. Axer, T. Bujnicki, Universitas, Kraków 2012, s. 555-570.

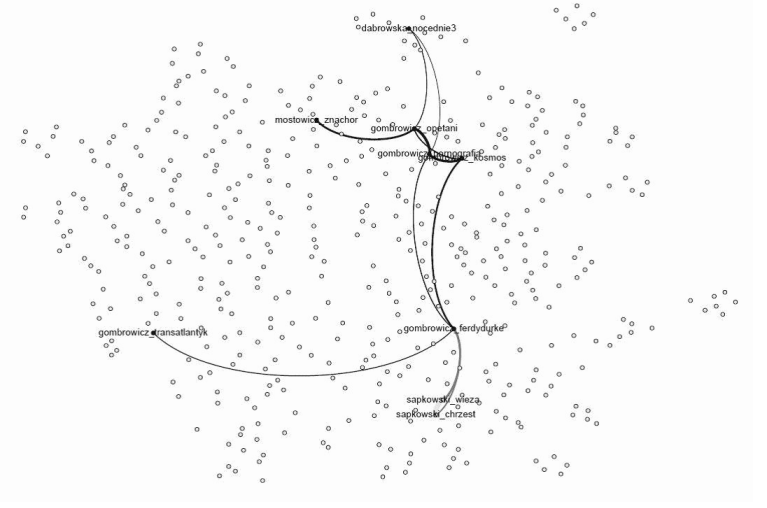
by sądzić, czy podobieństwo stylometryczne wynika ze wspólnoty rodzaju literackiego, czy z epoki literackiej. Nastoletnie powieści późniejszego autora *Nieboskiej* łączą się z romantyczną konkurencją dopiero za pośrednictwem pozytywisty – ale i poety – Gomulickiego. Bliskie sąsiedztwo do nich nie tylko *Ogniem i mieczem*, lecz (od drugiego końca) również XX-wiecznych powieści historycznych Orkana (*Kostka Napierski*), Reymonta (*Rok 1794*) i Kossak-Szczuckiej (*Przymierze*) może sugerować źródło ich archaizacyjnej stylizacji w twórczości trzech autorów wspomnianych poematów epickich. Warto zauważyć, że w lewym dolnym rogu wykresu – tego, w którym znalazły się teksty najstarsze i/lub powieści historyczne – dochodzi do lokalnego zakłócenia porządku chronologicznego.



Wykres 3. Wycinek sieci podobieństwa stylometrycznego; okolice polskich romantyków. Dla przejrzystości wykresu usunięto najsłabsze połączenia.

Ale najciekawszym zakłóceniem tego porządku – również powiązanym z Sienkiewiczem – jest jednak bliskie sąsiedztwo i silne połączenie między *Ogniem i mieczem* i *Trans-Atlantykiem*. Rzecz niby oczywista: i Sienkiewicz, i Gombrowicz mówią w tych okolicach „Paskiem", nawet jeżeli ten drugi czyni to w tonacji groteskowo-parodystycznej; a jednak nie należy zapominać, że na taki wynik nie miały wpływu ani szyk zdania, ani archaizowane elementy leksykalne (patrz Dodatki). Gombrowiczowski miecz parodii uderza zresztą

nie tylko w Sienkiewicza, lecz również w Ignacego Chodźkę i jego *Pamiętniki kwestarza*; efekt ten dziwić nie powinien, bo stylometryczne podobieństwo parodii literackiej i jej przedmiotu jest zjawiskiem znanym i dobrze opisanym[14]. W każdym razie *Trans-Atlantyk* tak skutecznie przyprawia sobie staropolską gębę, że dopływa najdalej na lewo (można powiedzieć, że na zachód!) wykresu ze wszystkich tekstów z drugiej połowy XX wieku. W ten sposób Gombrowicz jawi się jako autor stylometrycznie najbardziej proteuszowy z tych, których teksty znalazły się w omawianym korpusie. Jak widać na Wykresie 4, jego teksty – poza opisanym już spektakularnym rejsem – pojawiają się w trzech z czterech ćwiartek naszej sieci: *Ferdydurke* w prawej dolnej, *Opętani*, *Pornografia* i *Kosmos* po prawej u góry. Co ciekawe, sensacyjno-gotycki romans z 1939 roku wykazuje znaczące podobieństwo stylometryczne do *Znachora*. Czyżby dlatego, że jak Gombrowicz sam wyznaje, „zła literatura polska była dla mnie i ciekawa i pouczająca. Studiując [...] powieści Germana, Mniszkówny, Zarzyckiej, Mostowicza, odkrywałem rzeczywistość..."[15]. Już znacznie trudniej wytłumaczyć, dlaczego *Ferdydurke* łączy się z dwoma tekstami Sapkowskiego (*Chrzest ognia* i *Wieża jaskółki*). Albo skąd u tych samych *Opętanych* stylometryczne pokrewieństwo (prawda, że bardzo słabego stopnia) do *Nocy i dni*?

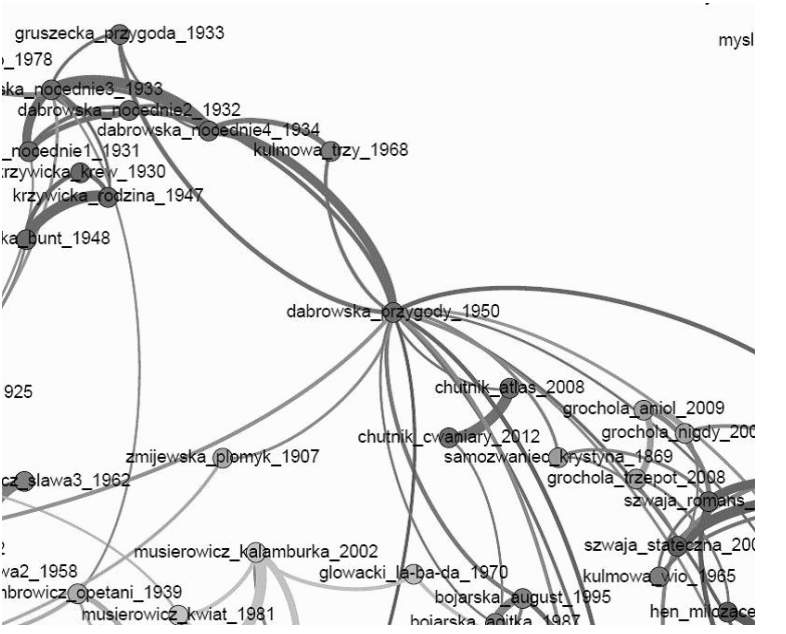


Wykres 4. Wycinek sieci podobieństwa stylometrycznego; autorski sygnał Gombrowicza. Dla przejrzystości wykresu usunięto wszystkie inne połączenia.

14 J. Burrows *Who wrote Shamela? Verifying the authorship of a parodic text*, „Literary and Linguistic Computing" 2005 nr 20(4), s. 437-450.

15 W. Gombrowicz *Dziennik 1953-1956*, Wydawnictwo Literackie, Kraków 1997, s. 108.

Skoro mowa o Marii Dąbrowskiej, to jeden z jej tekstów bije wszelkie rekordy pod względem liczby powiązań na wykresie. *Przygody człowieka myślącego* łączą się bowiem (Wykres 5) z aż czternastoma tekstami innego autorstwa. Czy dlatego, że *Przygody…* to rekonstrukcja nieukończonej powieści, połączenie gotowych rozdziałów z fragmentami brulionu autorki, pozbawione tak ważnej w twórczości Dąbrowskiej ostatniej redakcji autorskiej? Anna Kowalska pisze w swej przedmowie wręcz, że „to nie książka"…[16]

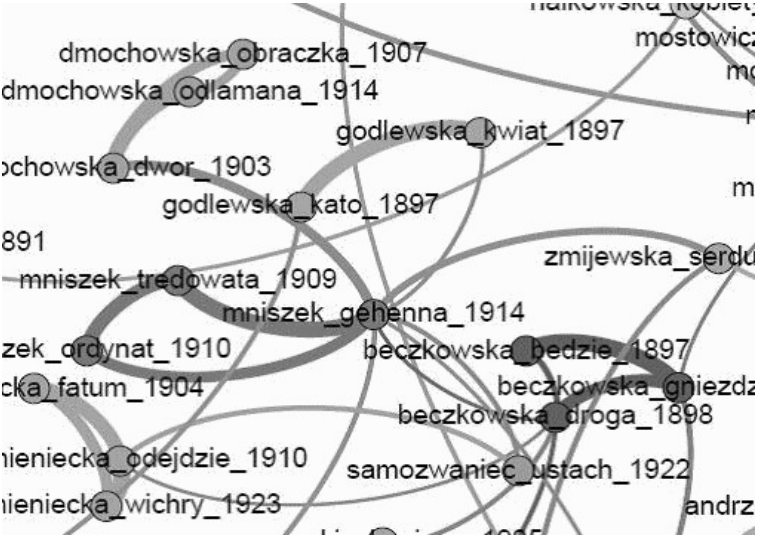


Wykres 5. Wycinek sieci podobieństwa stylometrycznego; ***Przygody człowieka myślącego***. Dla przejrzystości wykresu usunięto najsłabsze połączenia

Jest jeszcze jeden fragment opisywanej sieci, w którym szczególnie silne są powiązania między tekstami różnych autorów – a raczej, w tym przypadku, autorek. Chodzi o kilkanaście romansów z przełomu XIX i XX wieku autorstwa takich pisarek jak Mniszkówna, Grot-Bęczkowska, Dmochowska, Żmijewska czy Krzemieniecka. Węzłem łączącym stylometryczne cechy tej twórczości zdaje się (przynajmniej na podstawie Wykresu 6) *Gehenna*; ale równie mocno wiąże się z nimi parodia podgatunku, *Na ustach grzechu* Magdaleny Samozwaniec. I choć za główny przedmiot prześmiewczego dzieła córki

16 A. Kowalska, przedmowa do M. Dąbrowska *Przygody człowieka myślącego*, red. E. Korzeniewska, Czytelnik, Warszawa 1970, s. 9.

Wojciecha Kossaka przyjmuje się zazwyczaj *Trędowatą* (ta bowiem otwiera w dedykacji listę utworów sparodiowanych w *Na ustach grzechu*), to zapewne nie przypadek, że np. Grot-Bęczkowską wspomina Samozwaniec jako wzór własnych, nieudolnych prób literackich z dzieciństwa[17].



Wykres 6. Wycinek sieci podobieństwa stylometrycznego; ***Gehenna*** i ***Na ustach grzechu***. Dla przejrzystości wykresu usunięto najsłabsze połączenia.

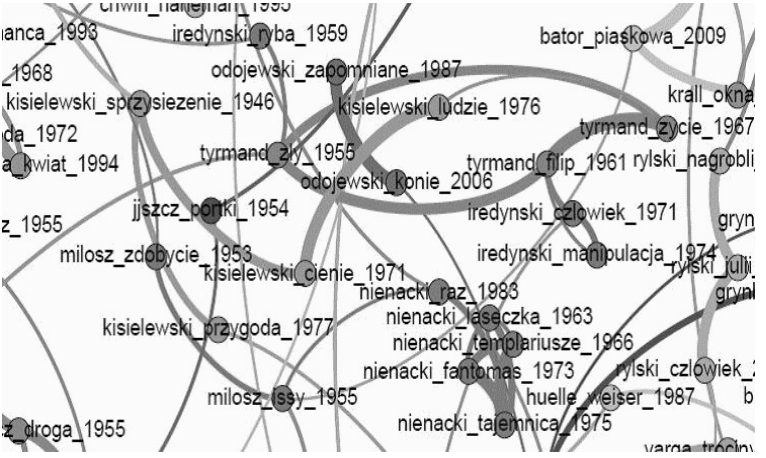
Innego rodzaju związek między autorami widać na Wykresie 7, gdzie silne podobieństwo łączy stylometrię trzech pisarzy: Kisielewskiego łączyła przecież bliska zażyłość z Tyrmandem („sporo myślałem sobie o Tyrmandzie. Lubiłem tego chłopca"[18]) i Iredyńskim („Iredyńskiego pijanego zgoła wczoraj wyrzuciłem na mordę"[19]). W tym samym wycinku sieci znalazły się też teksty Odojewskiego i Jana Józefa Szczepańskiego, ale, niestety, i Jan Józef Szczepański, jeszcze jeden z filarów pierwszego pokolenia *Tygodnika Powszechnego*, i Odojewski, o którym bardzo ciepło wyraża się w swych *Dziennikach* autor *Sprzysiężenia* („robi doskonałe wrażenie, podobno jest bardzo wybitny"[20]), są tutaj tylko „przelotem" – ich teksty (odpowiednio, *Portki Odysa*

17 M. Samozwaniec *Maria i Magdalena*, Wydawnictwo Literackie, Kraków 1960, s. 10.

18 S. Kisielewski *Dzienniki*, Iskry, Warszawa 2001, s. 197.

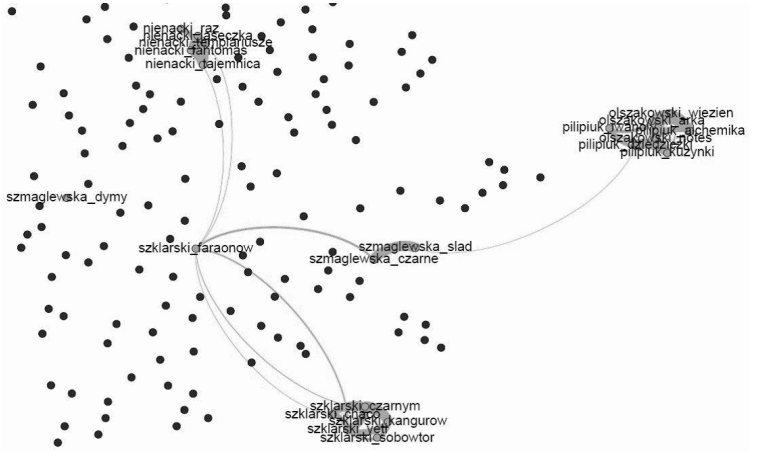
19 Tamże, s. 98.

20 Tamże, s. 568.



Wykres 7. Wycinek sieci podobieństwa stylometrycznego; Kisielewski, Tyrmand, Iredyński. Dla przejrzystości wykresu usunięto najsłabsze połączenia.

oraz *Zapomniane, nieuśmierzone…* i *I poniosły konie*) nie łączą się z „kisielową grupką" żadnymi silniejszymi więzami podobieństwa stylometrycznego. Całe szczęście, że PRL-owskie tajne służby nie dysponowały metodami stylometrycznymi i przez to nie mogły udowodnić autorowi *Ludzi w pieczarze*, że Kisielewski i Staliński to ta sama osoba – co w prezentowanej sieci staje się całkowicie oczywiste. A gdyby kiedykolwiek Andrzej Pilipiuk chciał się wypierać autorstwa kontynuacji przygód Pana Samochodzika (i tych, które popełnił pod własnym nazwiskiem, i tych, które podpisał jako Tomasz Olszakowski), wystarczy spojrzenie na naszą sieć (Wykres 8), by zobaczyć zwarte pipiliukowskie skupisko, jakże odległe od równie ciasno zbitego zlepku książek „prawdziwego" Nienackiego. Ten z kolei pisze dokładnie tak samo, gdy opiewa postać ORMO-wca-historyka sztuki, gdy pisze frywolny PRL-owski kryminał i gdy popełnia powieść czysto erotyczną w dekadenckich czasach schyłkowego komunizmu. Przygody jeszcze jednego idola PRL-owskiej młodzieży, Tomka Wilmowskiego, mają inną stylometrię, gdy opisuje je Alfred Szklarski (i zresztą powieść spoza cyklu podróżniczego, *Sobowtór profesora Rawy*, to istny sobowtór stylometryczny „Tomków"), a inną, gdy po śmierci tegoż kończy dzieło przyjaciel-ksiądz. Co ciekawe, Adam Zelga umieszcza swojego „Tomka" głęboko wewnątrz sieci literatury polskiej – a nie na stylometrycznym marginesie, na którym tkwią pierwowzory Szklarskiego. Na pewno z zupełnie innych powodów harcerskie powieści Seweryny Szmaglewskiej nie mają stylometrycznie absolutnie nic wspólnego z jej *Dymami nad Birkenau*; już bardziej przypominają pośmiertnego *Tomka w grobowcach faraonów* i pilipiukowskie „Samochodziki".



Wykres 8. Wycinek sieci podobieństwa stylometrycznego; powieści dla młodzieży, powieści dla dorosłych. Dla przejrzystości wykresu usunięto najsłabsze połączenia.

Pilipiuk i Szklarski nie są jedynymi pisarzami wypchniętymi na zewnętrzną orbitę naszej sieci i zresztą znaleźli się tam w całkiem niezłym towarzystwie. Mamy tu bowiem (Wykres 1) nazwiska takie jak Myśliwski, Stasiuk, Pilch i *last but not least*, Konwicki. Na pewno błędem byłoby konkludować, że w związku z tym wszyscy ci pisarze mają ze sobą coś wspólnego (bo przecież stylometrycznie nie łączy ich nic i właśnie dlatego zostali wypchnięci na peryferia wykresu!); trudno jednak oprzeć się wrażeniu, że pewna przewaga wśród tych outsiderów pisarzy powszechnie i współcześnie uznanych, może sugerować jakiś związek między oryginalnością (czy choćby tylko osobliwością) stylometryczną i popularnością czy uznaniem – bo przecież konkludowanie na tych podstawach o jakiejś „jakości" czy choćby tylko „sukcesie" literackim byłoby już grubym nadużyciem. Tak czy inaczej zapewne czas najwyższy na jakieś – może „spokojniejsze" –

Wnioski

Powstała więc mapa powiązań stylometrycznych między (dokładnie) 503 polskimi tekstami literackimi wybranymi z tego, co było dostępne, w sposób, który autorowi tego skromnego studium (zresztą angliście-komparatyście) wydawał się najbardziej reprezentatywny. Jestem przekonany, że kto inny sporządzający taką samą listę wymieniłby na inne nawet i połowę tych tekstów. Nie to jednak jest w tym studium najważniejsze, bo wielkie dyskusje o istnieniu i ewentualnym kształcie kanonu, jakie przetoczyły się przez literatury zachodnie, na razie trochę jakby ominęły literaturę

polską[21]. Ważne jest to, że niezależnie od tego, jacy autorzy i z jakich epok literackich znaleźli się w korpusie, metoda analizy i wizualizacji oparta na częstościach najczęstszych słów potrafi ułożyć teksty tych autorów i tych epok w pewnym porządku, a znaczna część odstępstw od tego porządku daje się wytłumaczyć w sposób niekłócący się znacznie z tym, co wiemy z historii i komparatystyki literackiej. Podobną prawidłowość zauważyłem zresztą w analogicznym korpusie 500 tekstów literatury anglojęzycznej – i podobnie jak w przypadku omawianego tu korpusu polskiego nie jest ona prostą funkcją historycznych zmian leksykalnych w języku (jeszcze raz przydał się ten jakże prosty a jakże skuteczny zabieg cullingu)[22].

Oczywiście wyniki te byłyby zapewne potraktowane znacznie poważniej, gdyby zostały uzyskane na cechach tekstu literackiego, do których jesteśmy bardziej przyzwyczajeni w praktyce interpretacji – gdyby nie opierały się na owych nieszczęsnych, pozbawionych kontekstu i w dodatku niezbyt „znaczących" słowach z górnych warstw listy rangowej. I – co należy dodać gwoli uczciwości – gdyby wiadomo było, dlaczego ciągi frekwencji słów najczęstszych tak skutecznie zdradzają autora tekstu. Jak pisze jeden z bardziej wpływowych stylometrów naszej doby, metody te „niebezpiecznie zakładają niezależność częstości jednych słów od drugich", choć przyznaje, że „sprawdzają się mimo tego podejrzanego założenia"[23]. Niewątpliwą zaletą stylometrii opartej na pojedynczych słowach jest przede wszystkim stosunkowa prostota metody i obliczeń; wszelkie próby automatycznego ustalania np. struktur zdaniowych mnożą problemy metodologiczne (szczególnie w języku tak silnie fleksyjnym, jakim jest polszczyzna); liczne próby dokonywane na n-gramach słownych wykazują znacznie mniejszą skuteczność niż na pojedynczych słowach[24]; skuteczny *parsing* semantyczny materiału literackiego to pieśń niedalekiej już, ale jednak przyszłości[25]. Na razie więc mapa literatury polskiej musi być taka, jaka jest; pierwsze mapy geograficzne też nie były bardzo dokładne i wykonane od razu najlepszymi metodami…

---

21 P. Wilczek *Czy istnieje kanon literatury polskiej?*, w: *Literatura polska w świecie. Tom I. Zagadnienia recepcji i odbioru*, red. R. Cudak, Gnome, Katowice 2006.

22 J. Rybicki *Visualizing literature: artistic statistics*, w: *The art of literature, art in literature*, red. B. Kucała, I. Curyło-Klag, M. Bleinert, Wydawnictwo UJ, Kraków 2014 (w druku).

23 S. Argamon *Interpreting Burrows's Delta: geometric and probabilistic foundations*, „Literary and Linguistic Computing" 2008 nr 23(2), s. 140.

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

25 Największe nadzieje wiąże piszący te słowa z działalnością Grupy Technologii Językowych (Politechnika Wrocławska), kierowanej przez dr. Macieja Piaseckiego.

Choć nie musimy wiedzieć, dlaczego papierek lakmusowy zmienia kolor pod wpływem kwasu czy zasady, a mimo to możemy stosować go na własnej skórze (przynajmniej w reklamach telewizyjnych) – trzeba przyznać, że rzeczywiście najdotkliwszą bolączką tego typu badań stylometrycznych jest brak teorii tłumaczącej w zadowalający sposób tak silną skuteczność stylometrycznego ustalania autorstwa (a więc podobieństwa między tekstami) na podstawie częstości najczęstszych słów. W fizyce „teoretycy" wskazują „doświadczalnikom", że powinni szukać bozonu Higgsa; komputerowy stylometra, czyli właśnie literaturoznawca „doświadczalny", takiego komfortu nie ma.

Na szczęście nie jest też całkowicie bezbronny względem materiału swych badań. Istnieje przecież potężna, tradycyjna wiedza literaturoznawcza, agregat interpretacji i historii literatury, wypracowana przez pokolenia badaczy literatury, która aż prosi się o zestawienie, zderzenie, pogodzenie lub zanegowanie właśnie w swoistych eksperymentach literaturoznawczych, ilustrowanych seriami stylometrycznych wykresów. Warto powtórzyć: konfrontacja takiego doświadczenia z istniejącą wiedzą o literaturze może nie wyjaśni, dlaczego najczęstsze słowa tak chętnie zdradzają autorstwo – ale jest nadzieja, że bazując na wielkich, nieobejmowalnych gołym okiem przez czytelnika czy interpretatora zbiorach tekstów, pozwoli wykryć nowe grupy tekstów, nowe rodzaje pokrewieństw, nowe interpretacje porównawcze. A to już nie byłoby takie obojętne dla „mainstreamowego" badacza literatury, na przykład (choć nie tylko) jej historyka i specjalisty od stylistyki[26]. Co więcej: współpraca takiego badacza ze stylometrą mogłaby pokazać – właśnie w obecnych warunkach uprawiania humanistyki – „że nie masz takowych terminów, z których by się *viribus unitis* przy boskich *auxiliach* podnieść nie można".

26 Wyzwania i nadzieje związane z taką współpracą zostały zasygnalizowane w pracy J. Rybicki i M. Heydel, *The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish*, „Literary and Linguistic Computing" 2013 nr 28(4), s. 708-717.

Dodatek 1:
Lista 827 najczęstszych słów występujących w 90% tekstów użytych w analizie (w malejącym porządku frekwencji).

1.	i	38.	tego	75.	jakby	112.	właśnie
2.	się	39.	może	76.	ani	113.	powiedział
3.	w	40.	ze	77.	potem	114.	mam
4.	nie	41.	dla	78.	które	115.	gdyby
5.	na	42.	gdy	79.	siebie	116.	niech
6.	z	43.	tu	80.	jednak	117.	im
7.	to	44.	ten	81.	gdzie	118.	która
8.	do	45.	nawet	82.	niego	119.	trzeba
9.	że	46.	pod	83.	sam	120.	wiem
10.	a	47.	ją	84.	oczy	121.	można
11.	jak	48.	pani	85.	ty	122.	były
12.	o	49.	ich	86.	lecz	123.	jako
13.	ale	50.	ma	87.	te	124.	wtedy
14.	co	51.	przed	88.	żeby	125.	mógł
15.	po	52.	teraz	89.	niej	126.	nagle
16.	tak	53.	on	90.	nas	127.	jestem
17.	za	54.	tam	91.	aby	128.	znowu
18.	już	55.	nic	92.	ona	129.	niż
19.	od	56.	który	93.	no	130.	panie
20.	jest	57.	przy	94.	raz	131.	więcej
21.	go	58.	była	95.	tych	132.	jeśli
22.	jej	59.	wszystko	96.	są	133.	ludzi
23.	tylko	60.	nim	97.	chwili	134.	wszyscy
24.	mnie	61.	tej	98.	dobrze	135.	wszystkie
25.	mu	62.	więc	99.	zawsze	136.	zaraz
26.	było	63.	by	100.	aż	137.	cię
27.	ja	64.	bardzo	101.	ta	138.	jeden
28.	jego	65.	kiedy	102.	nich	139.	zaś
29.	pan	66.	nad	103.	kto	140.	we
30.	bo	67.	bez	104.	pana	141.	sobą
31.	tym	68.	będzie	105.	je	142.	dalej
32.	jeszcze	69.	coś	106.	albo	143.	choć
33.	czy	70.	też	107.	domu	144.	której
34.	był	71.	u	108.	ku	145.	których
35.	mi	72.	być	109.	nigdy	146.	tę
36.	przez	73.	miał	110.	nią	147.	mój
37.	sobie	74.	ci	111.	przecież	148.	mówił

149.	lub	190.	rzeczy	231.	sposób	272.	wszystkim
150.	nikt	191.	dlaczego	232.	panu	273.	swoim
151.	drzwi	192.	czego	233.	chce	274.	jesteś
152.	człowiek	193.	rękę	234.	dzieci	275.	drugi
153.	ręce	194.	zupełnie	235.	swoją	276.	temu
154.	między	195.	znów	236.	masz	277.	głosem
155.	nam	196.	ciebie	237.	wśród	278.	całe
156.	twarz	197.	swoje	238.	głowy	279.	bóg
157.	miała	198.	dziś	239.	całą	280.	szybko
158.	ludzie	199.	zresztą	240.	stał	281.	jaki
159.	trochę	200.	ktoś	241.	wiele	282.	oto
160.	taki	201.	proszę	242.	ile	283.	swoich
161.	dopiero	202.	którą	243.	was	284.	stronę
162.	którego	203.	ziemi	244.	czasem	285.	później
163.	głowę	204.	pokoju	245.	powiedziała	286.	tymczasem
164.	cóż	205.	bardziej	246.	prawda	287.	razie
165.	chciał	206.	głos	247.	wiedział	288.	mówić
166.	także	207.	ojciec	248.	stary	289.	parę
167.	wszystkich	208.	my	249.	powiedzieć	290.	chociaż
168.	chwilę	209.	długo	250.	dni	291.	mimo
169.	czas	210.	twarzy	251.	człowieka	292.	tutaj
170.	chyba	211.	sama	252.	samo	293.	razy
171.	którym	212.	moje	253.	stało	294.	jedno
172.	czasu	213.	zaczął	254.	trzy	295.	razu
173.	lat	214.	głową	255.	obok	296.	rzecz
174.	słowa	215.	jakieś	256.	wiesz	297.	taka
175.	prawie	216.	nimi	257.	strony	298.	zapytał
176.	razem	217.	musi	258.	lepiej	299.	gdzieś
177.	dwa	218.	oni	259.	byli	300.	widział
178.	tyle	219.	jakie	260.	mną	301.	kiedyś
179.	takie	220.	swego	261.	dość	302.	naprawdę
180.	coraz	221.	którzy	262.	ojca	303.	świecie
181.	kilka	222.	mówi	263.	świat	304.	końcu
182.	cały	223.	dnia	264.	serce	305.	innych
183.	myśli	224.	dlatego	265.	matka	306.	dwie
184.	jakiś	225.	moja	266.	widać	307.	mają
185.	życie	226.	każdy	267.	myśl	308.	cicho
186.	dzień	227.	mogę	268.	nocy	309.	musiał
187.	wreszcie	228.	będę	269.	niby	310.	wam
188.	wie	229.	wcale	270.	mieć	311.	takiego
189.	życia	230.	pierwszy	271.	mogła	312.	takim

313. dwóch
314. góry
315. nieco
316. ręką
317. chcę
318. kobiety
319. jednego
320. swój
321. świata
322. noc
323. jednym
324. skąd
325. swej
326. miejsce
327. spojrzał
328. poza
329. oczach
330. dziecko
331. swojej
332. odpowiedział
333. pewno
334. serca
335. mało
336. roku
337. każdym
338. inaczej
339. natychmiast
340. nogi
341. mojej
342. śmierci
343. takich
344. kilku
345. drogi
346. pewnie
347. dużo
348. usta
349. mówiła
350. pół
351. jakoś
352. boże
353. mieli
354. miejscu
355. miejsca
356. głowie
357. muszę
358. taką
359. jednej
360. inne
361. będą
362. zbyt
363. mamy
364. innego
365. życiu
366. czasie
367. raczej
368. koło
369. jedna
370. byłem
371. wy
372. koniec
373. piersi
374. matki
375. cała
376. ciągle
377. duszy
378. zrobić
379. tą
380. drogę
381. wolno
382. zaczęła
383. idzie
384. jutro
385. pracy
386. samym
387. mniej
388. znaczy
389. oczu
390. miałem
391. wobec
392. śmierć
393. dał
394. chodzi
395. kobieta
396. jedną
397. wielki
398. miasta
399. często
400. drugiej
401. daleko
402. widzę
403. jaką
404. wody
405. chciała
406. wielkie
407. drodze
408. zapewne
409. powoli
410. czuł
411. ziemię
412. dom
413. wiadomo
414. górę
415. słowo
416. szczęście
417. wiedzieć
418. spokojnie
419. trudno
420. wyszedł
421. iść
422. miłość
423. siedział
424. moim
425. stanie
426. choćby
427. należy
428. moją
429. stała
430. myślał
431. tobie
432. nami
433. robić
434. jakąś
435. sprawy
436. samego
437. chcesz
438. mówię
439. jakże
440. całej
441. nikogo
442. został
443. właściwie
444. mówiąc
445. poszedł
446. patrząc
447. młody
448. szedł
449. zdawało
450. dać
451. włosy
452. podczas
453. wziął
454. pieniądze
455. da
456. woli
457. nasze
458. przynajmniej
459. dobry
460. stąd
461. zdaje
462. widok
463. niczego
464. słów
465. jakaś
466. nasz
467. głośno
468. kogo
469. dawno
470. słońce
471. okna
472. rzucił
473. zwykle
474. podniósł
475. one
476. będziesz

477. siły
478. mocno
479. prawo
480. ust
481. samej
482. lata
483. inny
484. czemu
485. źle
486. wielu
487. wielką
488. tuż
489. dobre
490. takiej
491. widzi
492. zamiast
493. krew
494. mąż
495. końca
496. mogą
497. chwila
498. stole
499. powiem
500. żadnego
501. inni
502. wszystkiego
503. uśmiechem
504. ręki
505. wszedł
506. myślę
507. mogło
508. wzrokiem
509. zrobił
510. jakim
511. żadnej
512. widzisz
513. stanął
514. całym
515. robi
516. znalazł
517. bądź
518. żadnych
519. początku
520. wrażenie
521. wprost
522. moich
523. godziny
524. byłoby
525. wzrok
526. nowe
527. ręku
528. rozumiem
529. sami
530. dzieje
531. rano
532. przyszedł
533. cztery
534. krwi
535. wczoraj
536. innym
537. byle
538. możesz
539. jaka
540. trzech
541. ciało
542. sprawę
543. powodu
544. wrócił
545. jakimś
546. mały
547. same
548. tobą
549. znam
550. wieczorem
551. niebo
552. przykład
553. dół
554. naszych
555. dobra
556. czegoś
557. imię
558. naszej
559. światło
560. pierwsze
561. stronie
562. ramionami
563. wszędzie
564. żyć
565. wielkim
566. przede
567. miało
568. jakiegoś
569. rąk
570. łzy
571. okno
572. pięć
573. spokój
574. czarne
575. żaden
576. drugą
577. rozmowy
578. drzewa
579. ciała
580. głębi
581. panem
582. lekko
583. mogli
584. ciężko
585. najbardziej
586. pierwszej
587. ściany
588. czoło
589. sprawa
590. naszego
591. wyraźnie
592. widząc
593. drugim
594. droga
595. światła
596. żal
597. twój
598. zwrócił
599. będziemy
600. łatwo
601. starego
602. wieczór
603. mieście
604. musiała
605. ostatni
606. pokój
607. jesteśmy
608. żyje
609. drugiego
610. przyjdzie
611. długie
612. milczeniu
613. stoi
614. prędzej
615. powietrze
616. stołu
617. kim
618. widzieć
619. wielkiej
620. kogoś
621. ramiona
622. jakiejś
623. prawa
624. powinien
625. żona
626. słychać
627. kieszeni
628. twoje
629. ramię
630. słyszał
631. słońca
632. postać
633. wieku
634. miały
635. sto
636. równie
637. rok
638. cisza
639. uwagi
640. różne

641. pełne
642. każdej
643. białe
644. oko
645. samą
646. znał
647. daj
648. pytanie
649. każdego
650. uwagę
651. znak
652. zęby
653. wielkiego
654. dłoni
655. zna
656. jemu
657. stały
658. roboty
659. powiedz
660. przyszło
661. wielka
662. prawdę
663. nowy
664. dziesięć
665. pewnego
666. stara
667. część
668. dzięki
669. otworzył
670. człowiekiem
671. głęboko
672. czekał
673. kroków
674. zatrzymał
675. mówią
676. tymi
677. ruch
678. stać
679. obu
680. myśleć
681. byłby
682. rękami
683. znaleźć
684. prosto
685. sił
686. celu
687. drugie
688. pamięci
689. małe
690. środku
691. udało
692. okiem
693. przyszła
694. dwadzieścia
695. powrotem
696. nieba
697. obraz
698. całego
699. oka
700. została
701. mógłby
702. głowa
703. powietrzu
704. sen
705. pierwszym
706. znać
707. ostatnie
708. wyszła
709. dawna
710. rady
711. szkoda
712. nikomu
713. dłużej
714. drogą
715. czeka
716. drzwiach
717. szukać
718. krótko
719. leżał
720. nowego
721. uśmiech
722. spać
723. naszym
724. twarzą
725. los
726. własne
727. wysoko
728. daleka
729. nasza
730. stół
731. własnej
732. najlepiej
733. odpowiedzi
734. zostać
735. ustach
736. powie
737. którymi
738. brak
739. każde
740. wygląda
741. ludźmi
742. rana
743. bliżej
744. pomocy
745. długi
746. wziąć
747. nóg
748. palce
749. któremu
750. możemy
751. pierwszego
752. robił
753. patrzeć
754. pewnym
755. niewiele
756. dziwne
757. godzin
758. każda
759. piękne
760. siedzi
761. trzymał
762. szeroko
763. stare
764. krok
765. słowem
766. boku
767. rękach
768. nadzieję
769. czekać
770. godzinę
771. porządku
772. komu
773. druga
774. złe
775. bok
776. chwilą
777. własną
778. myślą
779. podobne
780. nadziei
781. daje
782. nogami
783. złego
784. pamięć
785. wiedzą
786. wielkich
787. pierwsza
788. jakiej
789. późno
790. zostało
791. leży
792. poznać
793. drzwiami
794. stron
795. kolei
796. wodą
797. ruchu
798. trzeci
799. położył
800. zostanie
801. tacy
802. snu
803. stali
804. każdą

805.	wyjść	811.	dobrego	817.	własnym	823.	znajdzie
806.	jakich	812.	głosu	818.	niedawno	824.	wysoki
807.	muszą	813.	ostatniej	819.	żadna	825.	otwarte
808.	odpowiedź	814.	innej	820.	żadne	826.	nową
809.	mogły	815.	gorzej	821.	ostatnich	827.	ostatnim
810.	nigdzie	816.	nogach	822.	nowych		

Dodatek 2:
Lista 107 najczęstszych słów występujących we wszystkich tekstach użytych w analizie (w malejącym porządku frekwencji).

1.	i	28.	był	55.	coś	82.	ludzi
2.	się	29.	przez	56.	też	83.	wszyscy
3.	w	30.	sobie	57.	u	84.	wszystkie
4.	nie	31.	tego	58.	miał	85.	zaraz
5.	na	32.	może	59.	ani	86.	we
6.	z	33.	ze	60.	potem	87.	sobą
7.	to	34.	dla	61.	siebie	88.	tę
8.	do	35.	tu	62.	gdzie	89.	nikt
9.	że	36.	ten	63.	niego	90.	ręce
10.	a	37.	nawet	64.	sam	91.	taki
11.	jak	38.	pod	65.	oczy	92.	dopiero
12.	o	39.	ich	66.	te	93.	chciał
13.	ale	40.	ma	67.	ona	94.	czas
14.	co	41.	przed	68.	raz	95.	czasu
15.	po	42.	teraz	69.	tych	96.	lat
16.	tak	43.	on	70.	są	97.	razem
17.	za	44.	tam	71.	dobrze	98.	dwa
18.	już	45.	nic	72.	zawsze	99.	takie
19.	od	46.	który	73.	aż	100.	coraz
20.	jest	47.	przy	74.	ta	101.	cały
21.	go	48.	była	75.	nich	102.	dzień
22.	tylko	49.	wszystko	76.	kto	103.	długo
23.	było	50.	nim	77.	nigdy	104.	musi
24.	jego	51.	kiedy	78.	właśnie	105.	dni
25.	bo	52.	nad	79.	która	106.	trzy
26.	tym	53.	bez	80.	trzeba	107.	drugi
27.	czy	54.	będzie	81.	można		

Abstract

Jan Rybicki
JAGIELLONIAN UNIVERSITY (KRAKÓW)
First glimpse at a stylometric map of Polish literature

The author presents the analysis of circa 500 Polish literary texts, based on the analysis of frequency of usage of particular words and the visualisation of the outcome through network analysis. The result is discussed as a "map" of stylometric relationships between particular texts. The presence of two major "signals" was detected: auctorial and chronological. The most interesting exceptions were discussed in detail.