„Przekładaniec" nr 27/2013, s. 61–87
doi:10.4467/16891864PC.13.004.1286

JAN RYBICKI

STYLOMETRYCZNA NIEWIDZIALNOŚĆ TŁUMACZA

Abstract

Translator's Stylometric Invisibility

In a corpus of the writings of several authors, each author being represented by several texts, it is usually enough to compare the similarities between the frequencies of some 100 most frequent words (obviously, these usually include various function words rather than content words) in these texts to group the texts correctly by the authors. This paper investigates the phenomenon that translated texts also tend to be grouped by the original author rather than by the translator despite the fact that the most frequent words in a corpus of translations in no way maintain a one-to-one relationship with those in the original corpus. This is illustrated with examples of experiments performed on a variety of parallel sets of literary texts in English, French and Polish.

Key words: stylometry, most frequent words, cluster analysis, authorial signal, translatorial signal

Słowa klucze: stylometria, słowa najczęstsze, analiza skupień, sygnał autorski, sygnał tłumacza

Wprowadzenie

W danym zbiorze tekstów wystarczy znać częstości kilkuset (a czasem nawet kilkudziesięciu) najczęstszych słów, by pogrupować te teksty według ich autorów. A dokładniej: jeżeli weźmiemy po kilka tekstów (np. powieści) kilku różnych autorów, ciąg liczb dla danej książki (będący częstościami najczęstszych słów w całym tym zbiorze) będzie zwykle najbardziej podobny do ciągów częstości tych samych słów w innych dziełach tego sa-

mego autora. O takim autorskim „odcisku palca" wiadomo już co najmniej od czasu, gdy dwaj amerykańscy statystycy, Frederick Mosteller i David Wallace, policzyli słowa „funkcyjne" (czyli właśnie te najczęstsze: zaimki, przyimki, czasowniki modalne…) w esejach z serii „Federalist Papers", namawiających mieszkańców stanu Nowy Jork do ratyfikowania Konstytucji USA w latach 1787–1788. Teksty te pojawiały się anonimowo, ale autorstwo większości szybko odgadnięto i przypisano m.in. Alexandrowi Hamiltonowi i Jamesowi Madisonowi. Trudniejsze przypadki (12 tekstów) przypisali Madisonowi dopiero właśnie Mosteller i Wallace (1964). Od tego czasu atrybucja autorstwa oparta na częstościach najczęściej występujących słów stała się jedną z pewniejszych tzw. nietradycyjnych metod w tej dziedzinie i coraz częściej wykracza poza zwykłą atrybucję, szukając „sygnału" nie tylko autorskiego, lecz również gatunkowego czy chronologicznego, śladów redaktora, kopisty itp. (np. Dalen-Oskam, Zundert 2007; Eder, Rybicki 2013; Hoover 2004, 2007, 2012a; Jockers, Witten, Criddle 2008; Rybicki 2006; Rybicki, Eder 2011; Smith, Aldridge 2011).

Oczywiście przekonanie o istnieniu „ulubionych" słów autorów tekstów literackich (i innych) jest rozpowszechnione i zgodne z intuicją. Sęk w tym, że zwykle poszukuje się takich słów wśród leksemów mniej lub bardziej, ale jednak jakoś, „znaczących". David Lodge przedstawia w swym *Małym światku* postać wrednego lingwisty komputerowego Robina Dempseya, który doprowadza pisarza Ronalda Frobishera do niemocy twórczej, udowadniając mu statystycznie, że jego ulubionym słówkiem jest *grease*, „tłuszcz", lub pochodne *greasy*, „tłusty" lub „natłuszczony". Frobisher skarży się potem: „Nazajutrz, kiedy zasiadłem do biurka, by zabrać się pisania powieści, okazało się, że nie daję rady. Ilekroć potrzebny mi był przymiotnik, to zaraz «tłusty» przychodził mi do głowy" (Lodge 1992: 210).

Słowa najczęstsze to na ogół takie, których występowanie w tekście jest przez autora albo całkiem nieuświadomione (który pisarz liczy, ile razy użył spójnika *i*?), albo wymuszone przez gramatykę języka (każdy pisarz anglojęzyczny musi opatrzyć większość użytych rzeczowników którymś z dwóch przedimków, *a* lub *the*, choć i to nie w sposób losowy, tylko stosownie do kontekstu i znaczenia), albo przez mniej lub bardziej świadome zabiegi stylistyczne widoczne dopiero przy ogarnięciu całości tekstu. Sienkiewicz, na przykład, w każdej części swej trylogii używa częściej spójnika kontrastującego *ale*, ale pod koniec każdej z powieści coraz częściej pojawia się bardziej „poetyczna", czy też „patetyczna", alternatywa *lecz* (Rybicki 2010: 106–107).

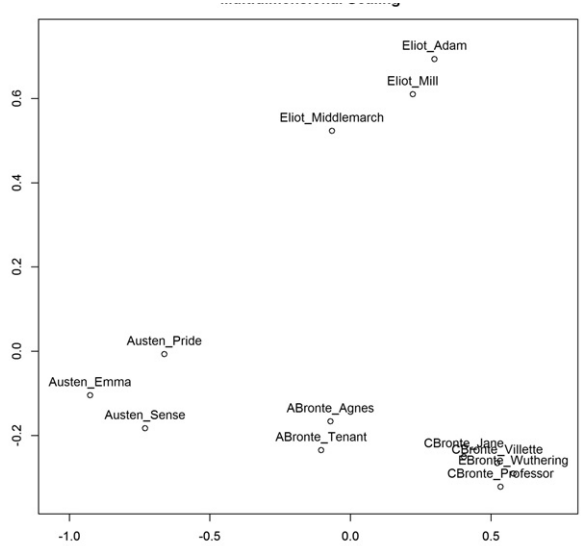
Jak pisze klasyk stylometrii John Burrows, „badając dzieła literackie, zachowujemy się tak, jak gdyby jednej trzeciej, dwóch piątych czy nawet połowy materiału po prostu w nich nie było" (1987: 1). I rzeczywiście: szukając w dziele Sienkiewicza „kochania" i „fantazji", kobiecej „gładkości" i męskiego „męstwa", „ojczyzny" i „przyjaźni" (czy choćby „waćpana" i „waszmości"), zwykle nie zwracamy uwagi na pojedyncze – choć znacznie częstsze – słowa „nieznaczące". A przecież pięćdziesiąt najczęstszych słów w *Ogniem i mieczem* to prawie 30% całości książki; sto najczęstszych słów pierwszej części trylogii to już dobrze ponad jedna trzecia powieści; za jej połowę odpowiada pierwsze pięćset *word types* (z prawie czterdziestu tysięcy). To występujące w każdym języku naturalnym zjawisko, znane jako (pierwsze) Prawo Zipfa, opisuje się w układzie współrzędnych hiperboliczną krzywą, która najpierw gwałtownie opada wzdłuż osi y (częstość słowa w tekście), po czym pod coraz mniejszym kątem dąży ku osi x (ranga słowa na liście frekwencyjnej). I choć krzywa ta ma nieco inny przebieg dla różnych języków, różnice między poszczególnymi tekstami w tym samym języku są najczęściej minimalne – ale najwyraźniej właśnie wystarczająco duże między tekstami tego samego autorstwa, by wykryto je analizą statystyczną. Warto dodać, że specjaliści nie ograniczają się do liczenia słów; równie popularna jest analiza n-gramów słownych (czyli zlepków n sąsiadujących słów) czy n-gramów literowych; mnożą się też konkretne metody obliczania „odległości" między ciągami częstości najczęściej występujących słów. A kiedy różnice metodologiczne dotyczą tak poważnych zagadnień jak (dla Anglosasów) autorstwo tekstów Szekspirowskich, spory bywają bardzo nieflegmatyczne (por. Hoover 2012b, w odpowiedzi na Vickers 2011, w odpowiedzi na Craig, Kinney 2009).

To wszystko przypomina – jak się za chwilę okaże, tylko do pewnego stopnia – sytuację w naukach doświadczalnych: dla konkretnego badanego korpusu wystarczy przeprowadzić powtarzalne doświadczenie, by uzyskać odpowiedź. Tylko że będzie to odpowiedź obowiązująca w tym jednym konkretnym zbiorze tekstów. By móc spodziewać się podobnej dla innych korpusów i, na przykład, w innych językach, trzeba wykonać bardzo dużo eksperymentów. Ale i wtedy trudno mówić o dowodzie – raczej o zgodności faktów empirycznych. Niestety: stylometria (może tylko na razie) różni się jednak od fizyki eksperymentalnej tym, że nie ma teoretycznego aspektu, który podpowiada, by (i jak, i gdzie) szukać bozonu Higgsa. Jeżeli więc na siłę trzymać się analogii między stylometrią a fizyką, to nie fizyką najnowszą; na razie jabłko dopiero spadło Newtonowi na głowę…

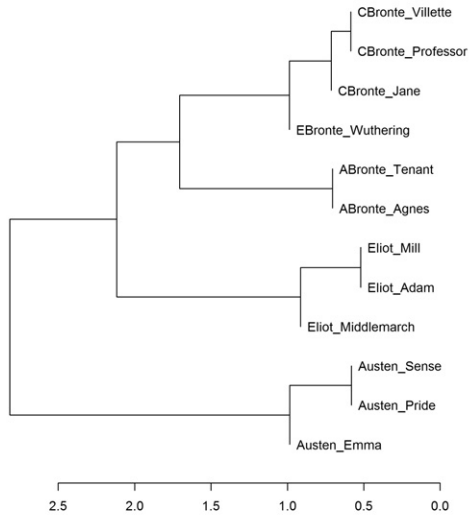
Stylometryczna analiza literacka – podstawy

Spadanie tego jabłka bada się w stylometrii na przykład tak: zlicza się częstości wszystkich słów w danym zbiorze tekstów i na tej podstawie sporządza listę słów najczęstszych. Potem zlicza się częstości tych słów w każdym z tekstów wchodzących w skład zbioru. Oczywiście teksty te są zwykle różnej długości, więc wartości trzeba znormalizować – najprostszą normalizacją jest podzielenie częstości każdego słowa w danym tekście przez rozmiar tego tekstu. Wytworzone w ten sposób ciągi częstości względnych są porównywane za pomocą jednej ze wspomnianych metod; najczęściej używana jest tzw. delta (Burrows 2002a). Im mniejsza wartość współczynnika delta między dwoma tekstami, tym większe jest ich stylometryczne podobieństwo, co w sytuacjach atrybucyjnych przekłada się na większe prawdopodobieństwo, że oba teksty wyszły spod jednego pióra. Dla większej jasności otrzymane wyniki można przedstawić na – z grubsza rzecz biorąc – dwóch rodzajach wykresów: „mapie" odległości (np. stosując procedurę zwaną skalowaniem wielowymiarowym, która „spłaszcza" wielowymiarową tabelę częstości do dwumiarowego wykresu, takiego jak na wykresie 1) albo wynikającego z tzw. analizy skupień „drzewka" podobieństw między tekstami, na którego najbliższe „gałązki" trafiają teksty o najpodobniejszych ciągach częstości słów (wykres 2). W obu przypadkach wykresy wykonano dla tego samego zbioru tekstów i takich samych parametrów: dla 100 najczęstszych słów, z usunięciem zaimków osobowych (co zwykle ułatwia atrybucję autorską w sytuacji, gdy narracja części tekstów występuje w pierwszej, a części w trzeciej osobie); poza tym brano pod uwagę tylko te słowa, które występują w 100% tekstów (*culling* na poziomie 100%).

W obu tych bardzo prostych i czytelnych przykładach wykresy dość skutecznie grupują teksty tych samych autorek. Na wykresie 1 utworzyły się wyraźne skupiska tekstów tego samego autorstwa; na wykresie 2 każda z pisarek też posiada własny „konarek" gromadzący „gałązki" jej tekstów. Równocześnie zwraca uwagę silne podobieństwo między twórczością trzech sióstr Brontë, przede wszystkim Charlotte i Emily; ich najbliższą sąsiadką jest trzecia mieszkanka rodzinnej parafii w Haworth, Anne. Można więc mówić, że oba wykresy ukazują – poza bardzo silnym sygnałem autorskim – również jakiś inny sygnał. Sygnał rodzinnego podobieństwa stylometrycznego (jeżeli istnieje coś takiego)? Wzajemnych wpływów literackich między trzema siostrami? A może „tylko" podobieństw wynika-



Wykres 1. Skalowanie wielowymiarowe 100 najczęstszych słów w 12 powieściach pisarek angielskich



Wykres 2. Analiza skupień 100 najczęstszych słów w 12 powieściach pisarek angielskich

jących z podobnej tematyki czy też chronologii powstawania ich utworów? Przy okazji warto wspomnieć, że plotki, jakoby powieści osób ukrywających się pod pseudonimami braci Bell, miały wyjść spod tylko jednej ręki (o co podejrzewano ich hulaszczego brata Patricka Branwella Brontëgo), choć historycznie i biograficznie nieprawdziwe, mają jednak jakieś podstawy (co najmniej) stylometryczne (Cooper Willis 1947).

W opisanych poniżej badaniach zastosowano skrypt autorstwa Macieja Edera, pisany dla środowiska programowania statystycznego R (Eder, Kestemont, Rybicki 2012). Skrypt przetwarza elektroniczne wersje tekstów i sporządza listy wszystkich słów użytych we wszystkich badanych tekstach wraz z ich częstościami występowania w poszczególnych tekstach; normalizuje te częstości tzw. Standaryzacją Z. Zgodnie ze wspomnianą już procedurą, skrypt pobiera do analizy słowa ze wskazanych przedziałów frekwencyjnych; wykonuje dodatkowe zabiegi poprawiające skuteczność atrybucji, takie jak usunięcie zaimków osobowych (szczególnie skuteczne w niefleksyjnych językach takich jak angielski) czy *culling* (automatyczne usunięcie słów zbyt charakterystycznych dla pojedynczych tekstów); porównuje wyniki dla poszczególnych tekstów; przeprowadza analizę skupień i przedstawia podobieństwa/odległości między tekstami na wykresie drzewkowym i/lub tworzy tzw. drzewko konsensusu bootstrapowego, czyli nowy wykres będący wypadkową wielu wykresów drzewkowych dla różnych wartości różnych parametrów (Baayen 2008: 157–160). Sporządzenie drzewka konsensusu umożliwia większą stabilność wyników, bo ten typ wykresu łączy dwa teksty tylko wtedy, gdy zostały one połączone konsekwentnie w określonym odsetku (np. połowie) „cząstkowych" wykresów analizy skupień (Eder, Rybicki 2011).

We wszystkich poniższych analizach użyłem tych samych wartości parametrów: zakres długości listy słów w każdej z analiz to 100–1000 najczęstszych słów w korpusie (zastosowanie inkrementu 10 sprawiało, że każdy wykres został oparty na konsensusie między wynikami dla 100, 110, 120 itd. słów). Analiza została ograniczona tylko do tych słów, które wystąpiły przynajmniej raz we wszystkich tekstach poddanych danej analizie (*culling* na poziomie 100%), co skutecznie eliminuje zbyt „łatwe" cechy słownictwa poszczególnych tekstów związane z ich tematyką i oczywiście nazwy własne, w tym imiona/nazwiska postaci. Teoretycznie można by się przecież obawiać, że wystąpienie w dwóch powieściach różnych autorów bohaterów o tym samym imieniu – jak choćby Jane Bennet w *Pride and Prejudice* i tytułowa protagonistka *Jane Eyre* – mogłoby doprowadzić do

omyłkowego przypisania obu powieści albo (jeszcze jednej!) Jane Austen, albo Charlotte Brontë. Zaimki osobowe zostały pominięte w analizie, ponieważ we wszystkich korpusach znalazły się teksty z narracją w pierwszej i trzeciej osobie, co szczególnie w powieściach angielskich i francuskich mogło zaburzać wyniki atrybucyjne.

W analizie częstości najczęstszych słów sygnał autorski zwykle brzmi tak silnie, że jest skutecznie wykrywalny także w większych korpusach. Jeżeli do użytego powyżej korpusu dodamy jeszcze kilkanaście czy nawet kilkadziesiąt innych tekstów innych pisarek i pisarzy, poprawność pogrupowania tekstów tych samych autorów jest nadal stuprocentowa. Ponieważ w tym przypadku zastosowano drzewka konsensusowe, wyniki prezentowane na obu wykresach są jeszcze bardziej wiarygodne (wykresy 3 i 4). Warto zauważyć, że wykres 3 zachowuje podobieństwo pisarstwa sióstr Brontë, że trzech autorów osiemnastowiecznych (Fielding, Richardson, Sterne) znajduje się na tej samej gałęzi, i że cały korpus dzieli się wyraźnie na dwie części. W tej „klasyfikacji" jedną grupę tworzą Dickens, Eliot, Hardy i siostry Brontë, drugą Austen, Trollope, Thackeray i wspomniani już pisarze wieku XVIII. Skuteczność atrybucji nie zmienia się po dodaniu kolejnych 30 tekstów i 6 autorów (Conrad, Galsworthy, Hardy, James,

Wykres 3. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 30 powieściach angielskich

Wykres 4. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 60 powieściach angielskich

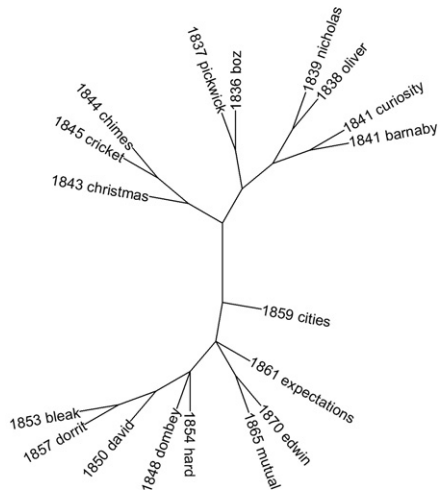
Kipling, Scott). Uwagę zwraca pokrewieństwo Conrada z Kiplingiem, Galsworthy'ego z Eliot oraz – po raz kolejny – sióstr Brontë.

Jak widać, siła sygnału autorskiego jest bardzo wyraźna w tak mierzonych różnicach i podobieństwach między tekstami. Niektóre z wymienionych wyżej połączeń między grupami poszczególnych autorów zdają się ponadto wskazywać na istnienie sygnału chronologicznego – choć nie dominuje on jasno we wszystkich większych relacjach międzyautorskich. Trudno byłoby udowodnić, na przykład, że istnienie wspólnej gałęzi tekstów Conrada i Kiplinga jest efektem osadzenia części tekstów obu pisarzy w sytuacjach kolonialnych; bardziej prawdopodobnym powodem tego podobieństwa jest właśnie podobny okres twórczości autorów *Jądra ciemności* i *Kima*. Z kolei całkiem niechronologiczny wydaje się związek między twórczością Jamesa i Trollope'a, Eliot i Galsworthy'ego.

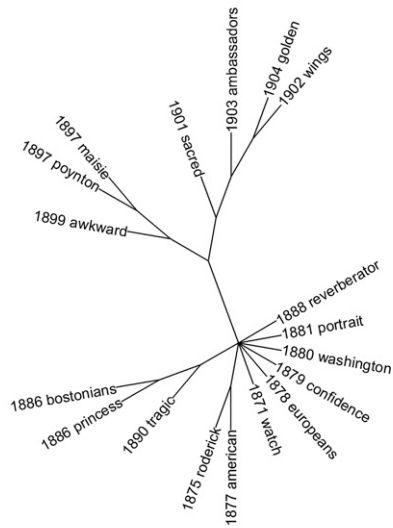
Skoro sygnał autorski zdaje się zagłuszać wszystkie inne, sygnału chronologicznego warto poszukać w twórczości jednego autora – najlepiej takiego, którego twórczość rozciąga się na wiele lat. Wykres 5 przedstawia najważniejsze dzieła Charlesa Dickensa wraz z datami pierwszych wydań. Co widać? Otóż analiza skupień przeprowadzona na ciągach czę-

stości najczęściej występujących słów „sama z siebie" (czyli bez udziału badacza) dzieli twórczość autora *Davida Copperfielda* na dwie połowy: w górnej części wykresu znalazły się dzieła wcześniejsze (1837–1845), w dolnej – teksty z późniejszych okresów twórczości (1848–1870). Warto zaobserwować, że – szczególnie wśród dzieł wcześniejszych – najbliżej siebie znalazły się utwory pisane w podobnych latach. I tak najbliższe sąsiedztwo *Sketches by Boz* i *Klubu Pickwicka* jest potwierdzeniem klasycznego podziału twórczości Dickensa, w którym teksty te reprezentują jego pierwszy, eksperymentalny okres (Strzetelski i in. 1987: 67). Wspólna gałąź *Nicholasa Nickleby'ego*, *Olivera Twista*, *Curiosity Shop* i *Barnaby'ego Rudge'a* to z kolei przejaw znacznej spójności stylometrycznej drugiego okresu. Nieco mniej uporządkowana jest dolna, późniejsza połowa wykresu, choć i tu zwraca uwagę odrębność ostatnich tekstów Dickensowskich: *Wielkich nadziei*, *Naszego wspólnego przyjaciela* i *Edwina Drooda*. Podobny wykres dla tej samej liczby powieści Henry'ego Jamesa (wykres 6) też dzieli twórczość tego pisarza na wczesną (dół) i późną (góra). Wbrew przyjmowanemu zwykle podziałowi twórczości autora *W kleszczach lęku* nie widać podziału na okres pierwszy (żartobliwie określany jako „James the First"; Guedalla 1920), zwieńczony *Portretem damy*, i drugi („James the Second"), rozpoczynający się *Bostończykami*. Z tej fazy twórczości na osobnej gałęzi znalazły się ostatnie trzy teksty, *The Spoils of Poynton*, *What Maisie Knew* i *The Awkward Age*; od tego zgrupowania odchodzi w bok najpierw przejściowe (a może tylko nietypowe) *The Sacred Fount*, a potem gałąź ostatniego okresu twórczości pisarza („The Old Pretender"), rozpoczętego *Ambasadorami*.

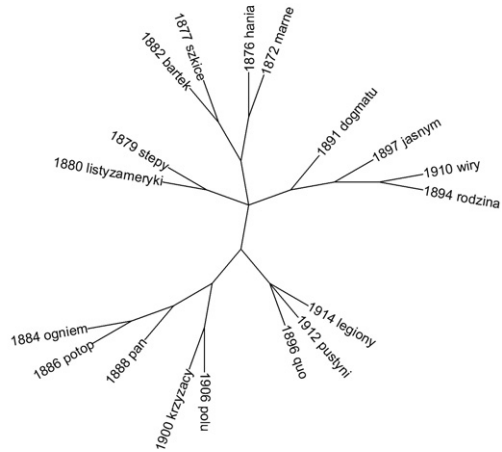
Nie zawsze jednak chronologia jest głównym sygnałem w twórczości jednego pisarza. Wykres 7 przedstawia układ podobieństwa stylometrycznego między powieściami i wybranymi nowelami Henryka Sienkiewicza. Ślady chronologii oczywiście są: wspólna gałąź łączy najwcześniejsze utwory pierwszego polskiego noblisty (młodzieńczą powieść *Na marne* i formy krótsze: *Hanię*, *Szkice węglem* i *Bartka zwycięzcę*). Jednak już sąsiadujące *Listy z podróży do Ameryki* i amerykańska nowela *Przez stepy*, mimo że są to pozycje wydane w odstępie zaledwie roku, zmuszają do zastanowienia się nad ewentualnym wpływem tematyki (choć oczywiście stuprocentowy *culling* skutecznie eliminuje z listy najczęstszych słów te, które mogłyby być związane ze znaczeniem poszczególnych tekstów). Z kolei prawa gałąź sugeruje podobieństwo między powieściami obyczajowymi: *Bez dogmatu*, *Rodziną Połanieckich* i *Wirami* a zbliżoną gatunkowo



Wykres 5. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 18 powieściach Charlesa Dickensa



Wykres 6. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 18 powieściach Henry'ego Jamesa.

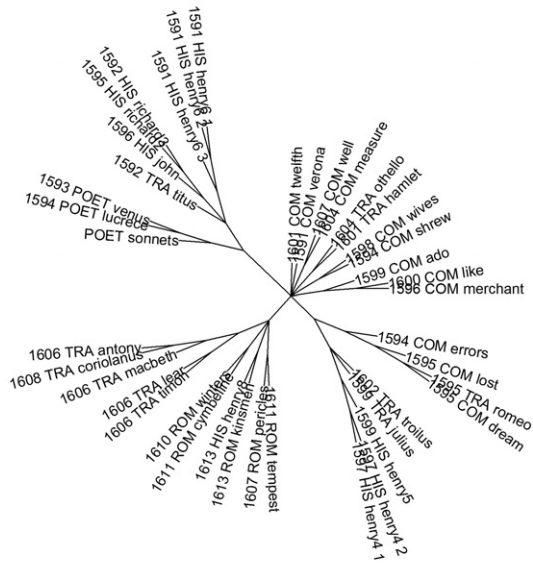


Wykres 7. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 18 powieściach i nowelach Henryka Sienkiewicza

Wykres 8. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 42 powieściach i nowelach polskich

i treściowo nowelą *Na jasnym brzegu* – tekstami, które trudno podejrzewać o przynależność do tego samego okresu twórczości Sienkiewicza. Cała dolna część wykresu jeszcze wzmacnia podejrzenia, że chodzi o sygnał gatunkowy, bo mamy tu powieści historyczne i przygodowe. Szczególnie ciekawe jest podobieństwo stylometryczne łączące *Quo vadis* z ostatnimi dziełami pisarza, *W pustyni i w puszczy* oraz niedokończonymi *Legionami* – jakby Sienkiewicz, zapewne nieświadomie, powracał po kilku mniej udanych powieściach do sprawdzonego idiomu swego największego światowego sukcesu. Różnice gatunkowe między powieściami historyczno-przygodowymi a obyczajowymi Sienkiewicza widać równie wyraźnie, gdy jego twórczość umieści się w kontekście prozy polskiej XIX i pierwszej połowy XX wieku (wykres 8). Co ciekawe, odrębności powieści historycznej nie widać ani u Prusa (*Faraon*), ani u Reymonta (*Rok 1974*), ani u Żeromskiego (*Popioły*). I zapewne dałoby się to wytłumaczyć bardziej konsekwentnym stosowaniem archaizacji języka *Ogniem i mieczem* czy *Krzyżaków*, gdyby nie – znów – stuprocentowy *culling*, który usunął z listy słów użytych w analizie wszelkie leksykalne elementy staropolszczyzny.

A więc w analizach stylometrycznych opartych na częstościach najczęściej występujących słów oprócz sygnału autorskiego i chronologicznego ujawnia się też sygnał gatunku? Ciekawe światło na to zagadnienie rzuca analiza związków gatunkowo-chronologicznych w kanonicznej twórczości Williama Szekspira – a więc w tych jego dziełach, których autorstwo zwykle przypisuje się Łabędziowi Awońskiemu. Wykres 9 przedstawia wyniki dla korpusu, którego poszczególne elementy wyznaczono według zwykłego podziału na pięć rodzajów/gatunków: utwory poetyckie, komedie, romanse, tragedie i historie (kroniki), dodatkowo opatrzone przypuszczalną datą powstania każdego utworu. Oczywiście w przypadku twórczości Szekspira wszelkie ustalenia chronologiczne są umowne, zaś kwestia autorstwa części sztuk i/lub ich fragmentów wciąż budzi kontrowersje, również (jak wspomniałem powyżej) w środowisku stylometrycznym. Przyjęta tu chronologia oparta jest (na wszelki wypadek) na najczęściej stosowanym układzie według Chambersa (1980); co więcej, przynajmniej jedna kategoria gatunkowa (*romances*) ma podstawy właśnie chronologiczne, skoro definiuje się ją m.in. jako „późne komedie". Należy również pamiętać, że *Sonety* były najprawdopodobniej pisane przez cały okres twórczości autora. Jak widać na wykresie, trudno mówić o „czystym" sygnale gatunkowym; raczej tylko o rodzajowym, skoro wszystkie dzieła poetyckie trzymają się wspólnej gałęzi. Poza tym dominuje jednak



Wykres 9. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w utworach poetyckich i dramatycznych Williama Szekspira

chronologia: jedyna gałąź obejmująca teksty tego samego gatunku to owe późne komedie, *romances*, z lat 1607–1613. Przy tym sąsiaduje ona z całą grupą tragedii z lat 1606–1608. Z kolei trzy z najwcześniejszych komedii z lat 1595–1596 (*Komedia omyłek*, *Stracone zachody miłości* i *Sen nocy letniej*) łączą się najbliżej z współczesną im tragedią o Romeo i Julii; wśród wczesnych historii (1591–1596) też znalazła się jedna współczesna im (ale i „problematyczna") tragedia, *Tymon Ateńczyk*. Co ciekawe, w tym samym obszarze wykresu tkwi też Szekspirowska poezja, a przecież *Wenus i Adonis* oraz *Gwałt na Lukrecji* pochodzą z tego samego, wczesnego etapu drogi twórczej Szekspira. Z uwagi na znaczną odległość dzielącą inne jego utwory od tych ostatnich trzech kategorii, można wręcz mówić o większym podobieństwie stylometrycznym zawartych w tej grupie utworów dramatycznych do poezji niż do innych dramatów. Najmniej uporządkowaną grupę tworzą komedie nieco późniejsze, ale jeszcze „nieromansowe", z lat 1594–1607, wraz z dwiema „wielkimi" tragediami (*Hamletem*, 1601, i *Otellem*, 1604) z tego samego okresu.

Stylometryczna analiza przekładu

Skoro częste słowa ujawniają – wśród tekstów wielu autorów – sygnał autorski, a w sprzyjających okolicznościach (na przykład przy ograniczeniu korpusu do twórczości jednego pisarza) sygnały chronologiczne i gatunkowe, przekładoznawcę stosującego takie metody szczególnie interesuje odpowiedź na pytanie, jaki sygnał okaże się najsilniejszy, gdy korpus zostanie zbudowany z tekstów tłumaczonych z jednego języka na drugi. Najbardziej oczywiste możliwości są dwie. Może dominuje sygnał tłumacza? W końcu słowa najczęstsze w języku przekładu to równocześnie te, które najtrudniej przypisać jednoznacznie ich odpowiednikom w języku oryginału… W dodatku do najczęstszych wśród najczęstszych należą np. angielskie *the* i *a*, które nie mają żadnych odpowiedników w języku polskim, i odwrotnie: w języku angielskim nie ma odpowiednika bardzo częstego polskiego *się*. A może jednak, w jakiś czarodziejski sposób, sygnał autora przebija się przez język tłumacza właśnie przez częstości tych najpospolitszych słów?

I znowu: można jedynie przeprowadzić szereg doświadczeń. Warto to zrobić, tym bardziej że wyniki innych badaczy nie wskazywały dotąd ani na jedną, ani na drugą z wyżej wymienionych możliwości, tylko trochę na jedną, a trochę na drugą. Oto bowiem przeprowadzone przez Johna Burrowsa porównanie przekładów twórczości łacińskiego poety Juwenalisa dokonanych przez dwóch znanych twórców angielskich: Samuela Johnsona i Johna Drydena, świadczy, że pierwszy z nich tłumaczy tak, jak pisał własne wiersze, drugi zaś prezentuje zupełnie dwa różne obrazy stylometryczne we własnej twórczości i w przekładach satyryka z Akwinum (Burrows 2002b). To sugeruje możliwość trzecią, kompromis w porównaniu z dwiema poprzednimi: w przekładzie miałby czasem dominować sygnał tłumacza (jak w przypadku Johnsona), czasem zaś – autora oryginału. Nie jest też wykluczone, że względna siła sygnałów autora i tłumacza może zależeć nie tylko od tłumacza, ale i od autora oryginału.

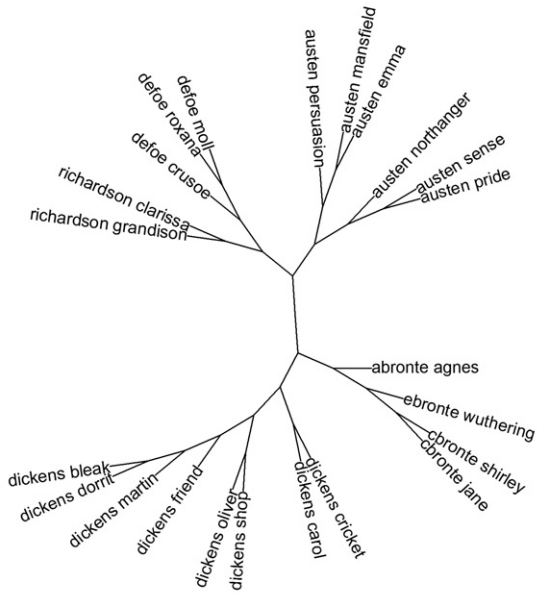
Tak jak w przypadku studium Burrowsa, dotychczasowe badania nad stylometrią przekładu koncentrowały się na tłumaczeniach jednego pisarza, bo wyeliminowanie różnic między „głosami" autorskimi pozwalało z większą pewnością na wychwycenie sygnału tłumacza – podobnie jak sygnał chronologiczny czy rodzajowo-gatunkowy ujawniał się w ukazanych powyżej przykładach (wykresy 5–9). W ten sposób poszukiwano więc śladów redaktorki angielskich przekładów Sienkiewicza (Rybicki

2011) i określono, w którym miejscu powieści jedna tłumaczka przejęła po drugiej przekład *Nocy i dnia* Virginii Woolf (Rybicki, Heydel 2013).

Nie należy zapominać o jeszcze jednej zmiennej w tym skomplikowanym równaniu: o językach, między którymi dokonuje się akt przekładu. Tak jak nie ma gwarancji, że wszystkie korpusy w tym samym języku będą zachowywać się podobnie, różnice w zjawiskach stylometrycznych mogą wystąpić między poszczególnymi parami języków i/lub między przekładem z języka A na język B a przekładem z języka B na język A. Trochę (ale tylko trochę) więcej pewności może dać porównanie drzewek bootstrapu konsensusowego dla tekstów oryginalnych z tymi, które sporządzi się dla ich przekładów – a jeszcze lepiej, jeżeli będą to przekłady na więcej niż jeden język.

Właśnie dlatego zebrałem dwa zasadnicze korpusy: 23 klasyczne powieści angielskie (siostry Brontë, Jane Austen, Daniel Defoe, Charles Dickens, Samuel Richardson) i 48 równie klasycznych powieści francuskich (Balzac, Dumas ojciec, Dumas syn, Flaubert, France, Hugo, Proust, Sue, Verne, Zola). Ponadto dla obu zestawiłem „korpus cieni", czyli ich polskich przekładów, odpowiednio 24 i 50 tekstów. Wreszcie trzecim elementem doświadczenia były korpusy angielskich przekładów tych samych tekstów francuskich oraz francuskich przekładów opisanego powyżej korpusu angielskiego, odpowiednio 50 i 29 (nieznaczne różnice w liczbie tekstów wynikają z istnienia więcej niż jednego przekładu niektórych powieści). Oczywiście znacznym ograniczeniem okazała się dostępność tekstów elektronicznych w dwóch językach – a i tak musiałem zrezygnować z polskich przekładów Richardsona, których nie odnotowują katalogi ani Biblioteki Jagiellońskiej, ani Narodowej. Nie udało mi się też zebrać równie zróżnicowanych korpusów angielskich i francuskich przekładów literatury polskiej, bo te dostępne w legalnych i nielegalnych wersjach elektronicznych ograniczone są w zasadzie tylko do twórczości Sienkiewicza i Lema. Znaczna dysproporcja w liczbie tekstów między korpusem oryginalnie angielskim a oryginalnie francuskim wynika z kolei ze znacznie większej dostępności przekładów literatury francuskiej niż angielskiej na język polski, co wydaje się dość ciekawym przeżytkiem jeszcze z wieku XIX, kiedy recepcja pisarzy francuskich była u nas znacznie obszerniejsza niż recepcja pisarzy angielskich.

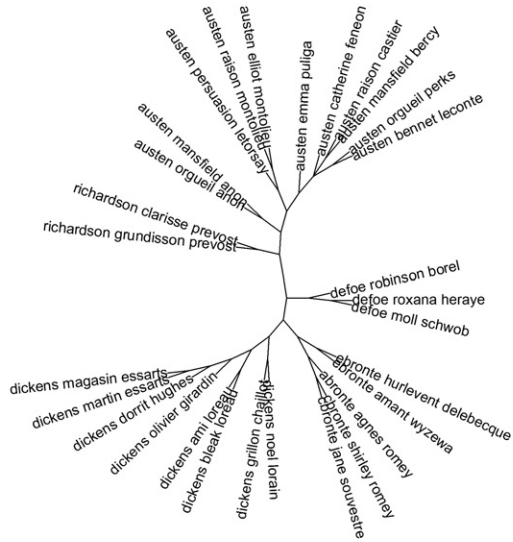
Wykres 10, sporządzony dla korpusu angielskich oryginałów, nie odbiega od wcześniej prezentowanych wyników dla tekstów oryginalnych: sygnał autorski jest wykrywalny w 100%, ale sygnału chronologicznego



Wykres 10. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 23 powieściach angielskich.

również nie należy bagatelizować, skoro wykres dzieli się na powieści wcześniejsze (góra) i późniejsze (dół). W dolnej części widać też po raz kolejny podobieństwo pisarstwa sióstr Brontë. Co dzieje się po „przetłumaczeniu" tego korpusu na język francuski? Właściwie niewiele: starsze teksty wciąż grupują się u góry wykresu, nowsze w dole; co więcej, mimo czworga różnych tłumaczy, pięć francuskich wersji powieści sióstr z plebanii w Haworth nadal tworzy własną gałąź (Wykres 11). Warto przy tym zauważyć, że *Shirley* Charlotte Brontë przetłumaczona przez Charlesa Romeya i A. Rolet (pełnego imienia – a więc i płci – A. Rolet nie udało się odszyfrować) najpierw łączy się z *Jane Eyre* tej samej autorki przetłumaczoną przez Noëmi Lesbazeilles Souvestre, a dopiero potem z *Agnes Grey* siostry Anne w przekładzie tandemu Romey–Rolet – co stanowi bardzo ładny przykład dominacji głosu autorek nad (wspólnym?) głosem tłumaczy.

Główną różnicą między wykresami dla oryginałów i dla francuskich przekładów jest rozdzielenie wspólnej gałęzi Defoe (który mimo trzech różnych tłumaczy posiada jednak własną gałąź) i Richardsona serii *Lettres*



Wykres 11. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 29 francuskich przekładach powieści angielskich z wykresu 10

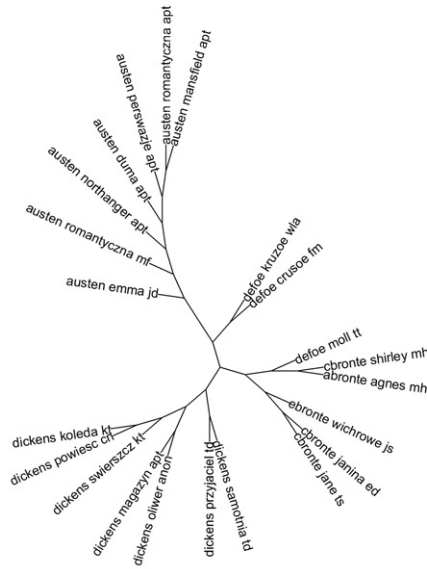
angloises [*sic*], bo tak powiązał *Clarisse* i *Grundisson* [*sic*] ich tłumacz *Abbé* Antoine Prévost, lepiej znany jako autor *Manon Lescaux*. Podobnie jak na wykresie 10, francuski Richardson (choć pozbawiony towarzystwa autora *Robinsona Crusoe*) sąsiaduje z powieściami Jane Austen, najbliżej zaś z pierwszymi, znacznie skróconymi i anonimowymi przekładami *Dumy i uprzedzenia* i *Mansfield Parku*, wydanymi jeszcze za życia (wówczas wciąż anonimowej) autorki. Na próżno jednak dopatrywać się ich anonimowego tłumacza w sławnym, choć dość awanturniczym proboszczu – zmarł na wylew podczas leśnej przechadzki w roku 1763, czyli na dwanaście lat przed narodzinami Jane Austen (z kolei córki innego proboszcza).

Jeżeli w korpusie przekładów francuskich znalazły się dwa tłumaczenia tej samej książki, zwykle stanowią swoje najbliższe sąsiedztwo. Byłoby to może bardziej oczywiste, gdyby analizę przeprowadzono na liście frekwencyjnej zawierającej jakiekolwiek słowa charakterystyczne dla treści, temu jednak zapobiega *culling* na poziomie 100%. Od tej reguły są zresztą interesujące wyjątki. Wspomniany już anonimowy skrót *Pride and Prejudice – Orgueil et Préjugé* z 1813 roku – leży daleko od innych przekładów

tej powieści Jane Austen, które z kolei jak najbardziej z sobą sąsiadują: *Orgueil et prévention* Éloïse Perks z 1822 roku i rozkosznie zatytułowanych *Les Cinq Filles de Mrs Bennet*, przetłumaczonych wspólnie przez Valentine Leconte i Charlotte Pressoir ponad sto lat później (1932). Podobnie zachowuje się drugi skrócony anonim i bezpośredni sąsiad pierwszego, *Mansfield-Park* (1815), położony całkiem daleko od *Mansfield Park* Léonarda Bercy'ego z roku 1945. Natomiast sygnał tłumacza (i/lub sygnał chronologiczny) widać w przypadku dwóch starych przekładów autorstwa szwajcarskiej pisarki Isabelle de Montolieu, bo jej *Persuasion*, czyli *La Famille Elliot, ou L'Ancienne Inclination* (1821), łączy się najpierw z jej *Pride and Prejudice*, czyli *Raison et Sensibilité, ou Les Deux Manières d'aimer* (1815), a dopiero potem z *Persuasion*, przełożoną kilkadziesiąt lat później przez „Madame Letorsay" (1882).

W porównaniu z korpusem przekładów francuskich polskie tłumaczenia (wśród których, jak już wspomniałem, zabrakło Richardsona) zachowują się dość podobnie, choć z jedną fałszywą atrybucją (wykres 12). Jak widać, grupa sióstr Brontë i tu pozostała nietknięta, ale teraz dołącza do nich *Dola i niedola sławnej Moll Flanders* Teresy Tatarkiewicz; warto zauważyć bezpośrednie sąsiedztwo obu polskich przekładów *Jane Eyre*, choć najwcześniejszy polski przekład, *Janina* Emilii Dobrzańskiej (1881), to wersja skrócona i zapewne dokonana za pośrednictwem przekładu francuskiego (czyżby autorstwa Souvestre). Poza tym jednak dominuje sygnał autorski. Wśród polskich Dickensów widać silne podobieństwo dwóch wersji *A Christmas Carol*: Krystyny Tarnowskiej *Kolędy prozą czyli opowieści wigilijnej o duchu* oraz *Wigilji Bożego Narodzenia* Cecylii Niewiadomskiej; ta choinkowa gałązka dopiero potem łączy się z kolejnym przekładem Tarnowskiej (i kolejnym świątecznym Dickensem), *Świerszczem za kominem*. Podobne stylometrycznie są dwa przekłady Tadeusza Dehnela (*Samotnia* i *Nasz wspólny przyjaciel*). Bliskie sąsiedztwo *Magazynu osobliwości* Anny Przedpełskiej-Trzeciakowskiej i anonimowego *Oliwera Twista* nie oznacza, że wyszły spod tego samego pióra, bo *Oliwer* to przekład jeszcze dziewiętnastowieczny. Z kolei sama Anna Przedpełska-Trzeciakowska przedstawia się zupełnie inaczej jako tłumaczka Dickensa i jako tłumaczka Jane Austen, jednak jej przekłady autorki *Emmy* (jedynej powieści Austenowskiej, której nie tłumaczyła) trzymają się razem, nierozdzielone innymi spolszczeniami Austen.

Literatura francuska w oryginale daje się równie łatwo pogrupować pod kątem sygnału autorskiego jak literatura angielska, bo, jak widać na



Wykres 12. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 24 polskich przekładach powieści angielskich z wykresu 10.

wykresie 13, każdy z tekstów ujętych w tym korpusie sąsiaduje najpierw z utworami tego samego autora. Najbardziej odstaje od innych pisarzy Jules Verne, co może być wynikiem odrębności gatunkowej nie tylko względem (zbliżonych do siebie) Flauberta czy Zoli, ale i względem nieco innej powieści przygodowej, reprezentowanej przez większe skupisko ojca i syna Dumasów wraz z Eugène Sue. Co ciekawe, jedyna powieść Dumasa juniora łączy się najpierw z tasiemcami Sue, a dopiero potem z twórczością protoplasty. Wyraźne (choć może mniej zrozumiałe) jest stylometryczne podobieństwo między tekstami France'a, Hugo, Prousta i (do pewnego stopnia) Balzaka.

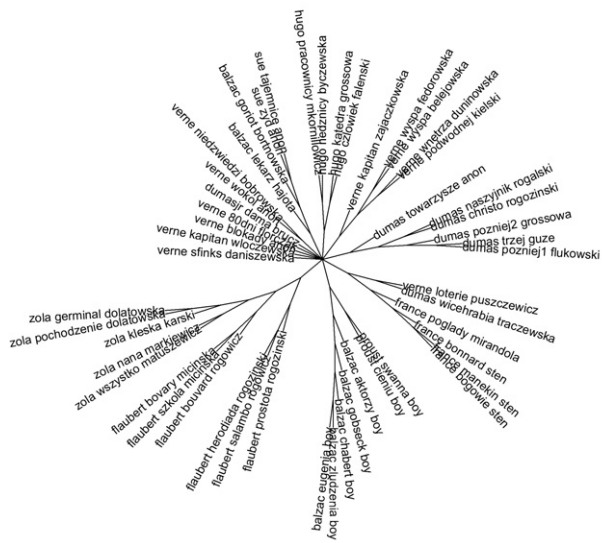
Część z tych podobieństw została ocalona w tłumaczeniu angielskim, które wykazuje poza tym daleko idącą wierność sygnałowi autorskiemu. Verne wciąż stanowi wyraźnie odrębne skupisko, Flaubert znów łączy się z Zolą. Natomiast znacznie mniej dają o sobie znać podobieństwa stylometryczne między pozostałymi pisarzami: jak widać na wykresie 14, każdy z nich posiada teraz własną gałąź wychodzącą ze wspólnego pnia wykresu. Wyjątkiem jest tylko mylne dołączenie *Zbrodni Sylwestra Bonnard*

Wykres 13. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 48 powieściach francuskich

Wykres 14. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 51 angielskich przekładach powieści francuskich z wykresu 13

Anatole'a France'a do obu tekstów Sue, w czym trudno dopatrywać się podobieństw chronologicznych (przekład Lafcadia Hearna ukazał się w roku 1890, a angielskie wersje obu romansów Sue pochodzą z wydania dzieł zebranych z 1912); w dodatku trudno podejrzewać Hearna (lepiej znanego jako Koizumi) o autorstwo przekładów *Tajemnic Paryża* i *Żyda wiecznego tułacza*, skoro ten oryginalny pisarz grecko-irlandzko-amerykańsko-japoński zmarł w Tokio w roku 1904. Odrębność stylometryczną powiększa też w angielskim przekładzie *Dama kameliowa* Dumasa syna, przetłumaczona – jak wiele tekstów w tym korpusie – przez jeszcze jednego anonima.

Znacznie mniej nieustalonych z nazwiska tłumaczy znaleźć można w korpusie polskich przekładów literatury francuskiej. Cóż z tego, skoro i tak w spolszczeniach tych najbardziej gubi się sygnał autorski (wykres 15). Co prawda Zola oraz Flaubert zachowują i własną odrębność, i swe wzajemne podobieństwo, które można było zaobserwować w korpusach oryginalnym i angielskim; Wiktor Hugo to Victor Hugo; Anatol France to Anatole France. Za to polski Dumas ojciec gubi gdzieś *Wicehrabiego de Bragelonne*, a Verne rozpada się na kilka różnych gałęzi. Balzac po polsku ma dwa oblicza: Boyowskie i nie-Boyowskie. Co więcej, Balzac



Wykres 15. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 50 polskich przekładach powieści francuskich z wykresu 13

Wykres 16. Drzewko bootstrapu konsensusowego 100–1000 najczęstszych słów w 50 polskich przekładach powieści francuskich z wykresu 13, uzupełnionych o 3 oryginalne teksty Tadeusza Boya Żeleńskiego

Tadeusza Żeleńskiego to najbliższy sąsiad Prousta, też przetłumaczonego przez autora *Słówek* – jest to chyba najwyraźniejszy (i na pewno jedyny tak oczywisty) w całej tej serii doświadczeń efekt stylometrycznego sygnału tłumacza. I nie tylko tłumacza: jak pokazuje wykres 16, teksty Boya tworzą zwartą grupę nawet wtedy, gdy do korpusu doda się jego własne – jakże gatunkowo odmienne – eseistykę i wspomnienia.

Uwagi końcowe

> Tekst przekładu (…) jest uznawany za zadowalający przez większość wydawców, recenzentów i czytelników, gdy płynnie się czyta, gdy przez brak jakichkolwiek własnych cech językowych czy stylistycznych wydaje się on przeźroczysty, sprawiając wrażenie, że odzwierciedla osobowość lub intencję obcojęzycznego pisarza albo esencjonalne znaczenie obcojęzycznego tekstu – wrażenie, że przekład wcale nie jest przekładem, lecz „oryginałem" (Venuti 1995, 1).

Lawrence Venuti chyba nie spodziewał się – ubolewając nad powszechnym wśród konsumentów apetytem na niewidzialność tłumacza (czy wręcz nad powszechnym przekonaniem o tej niewidzialności) – że jeszcze jedna jej forma zostanie kiedyś ujawniona przez statystyczne badania częstości słownictwa. Tymczasem większość prezentowanych powyżej wykresów dobitnie niewidzialność tę potwierdza, minimalizując – przynajmniej stylometrycznie – ślad tłumacza w tekście przekładu. Słaba to pociecha, że może stylometria jakby minimalizuje też znaczenie wpadek przekładowych.

Podstawowym problemem tego typu badań – które starałem się bardzo konsekwentnie nazywać stylometrycznymi, nie zaś stylistycznymi (czy choćby komputerowo-stylistycznymi) – jest przede wszystkim brak teoretycznych podstaw samego zjawiska tak skutecznej atrybucji autorskiej w korpusach tekstów oryginalnych. Jak pisze jeden z statystyków zajmujących się tą problematyką, metody te „niebezpiecznie zakładają niezależność częstości jednych słów od drugich", a jednak „działają lepiej niż cokolwiek innego mimo tego podejrzanego założenia" (Argamon 2008: 140). Tymczasem częstości najczęstszych słów okazują się wystarczającym kryterium rozpoznawczym indywidualnych cech pisarstwa (których wolę nadal nie nazywać stylem); co więcej, kryteria te mogą stać się nowym i obiektywnym narzędziem klasyfikacji i samych tekstów literackich, i okresów twórczości pisarzy, i wreszcie całych grup autorów.

Jeżeli jednak przyjmie się za dobrą monetę rosnącą liczbę doświadczalnych potwierdzeń tego zjawiska w zwykłej atrybucji autorskiej, paradoksalnie ów zasadniczy problem metodologiczny znika w stosujących te same metody badaniach nad tekstem w przekładzie. Skoro bowiem metoda działa dla tekstów w oryginale – niezależnie od tego, jak to się dzieje – zastosowanie jej do porównywania przekładów jest już tylko prostą konsekwencją. W końcu pracownicy laboratoriów biochemicznych zapewne nie wiedzą dokładnie, jak to się dzieje, że papierek lakmusowy barwi się na czerwono pod wpływem kwasów, a na niebiesko pod wpływem zasad.

Pozostaje jednak zasadnicze pytanie, jak to się dzieje, że czasem wystarczy choćby 100 najczęstszych słów w danym korpusie, by na podstawie ich znormalizowanej częstości poprawnie pogrupować teksty napisane przez kilku różnych autorów (co jest już samo w sobie zagadką), nawet jeżeli badamy nie ich oryginały, lecz przekłady sporządzone przez kilku różnych tłumaczy. Skąd się bierze niemal doskonałe odwzorowanie układu

z wykresu 9 (powieści angielskie, oryginały) na wykresie 11 (ich polskie przekłady), jeżeli wiadomo, że jest ono wynikiem badania względnych częstości tych konkretnych słów angielskich:

> the, to, and, of, a, in, that, it, was, as, not, with, be, for, had, but, have, is, at, so, on, this, by, all, said, mr, if, would, what, will, which, from, no, or, were, one, very, when, been, an, are, could, do, there, upon, more, who, out, am, now, sir, man, should, such, than, little, any, has, know, up, then, into, miss, much, how, say, must, own, some, time, well, good, think, see, can, did, may, dear, before, never, shall, other, only, about, too, being, might, made, come, make, again, down, like, yet, after, great, thought, every, two, go,

i tych konkretnych słów polskich:

> i, się, w, na, z, że, do, to, a, o, jak, tak, pan, co, ale, jest, po, za, czy, by, tym, od, tego, bardzo, tylko, sobie, dla, już, było, przez, był, gdy, kiedy, może, pana, była, jeszcze, ze, teraz, jednak, więc, nic, który, gdyby, ten, być, domu, będzie, wszystko, ani, przy, niż, tej, lecz, nawet, przed, nigdy, bez, tam, też, bo, które, pod, tu, jestem, iż, proszę, mam, coś, która, nad, jeśli, zawsze, gdzie, panie, dobrze, siebie, raz, chwili, miał, są, choć, więcej, aby, te, bardziej, miała, powiedział, mogła, właśnie, sam, sposób, wiem, tych, gdyż, u, wiele, słowa, ta, można.

Poszukując na tych dwóch listach jednoznacznych odpowiedników, łatwo zauważyć, że mamy tu do czynienia z różnym stopniem braku odwzorowania. Po pierwsze mamy tu do czynienia z istnym zatrzęsieniem przyimków – te jednak wchodzą w obu językach w tak różnego rodzaju i znaczenia związki z innymi częściami mowy (osławione „frejzale", których tak boją się studenci anglistyki), że trudno tu mówić o jakiejkolwiek bezpośredniej ekwiwalencji. Po drugie, zwroty grzecznościowe, spośród których z angielskich *mister, sir* i *miss* ocalały w tłumaczeniu tylko *pana* i dwuznaczne *panie* (wołacz od *pan* i liczba mnoga od *pani*). Warto zauważyć, że fleksyjność języka polskiego ma znaczenie nawet w tak specyficznej liście, skoro angielskiemu *said* odpowiada co prawda *powiedział*, ale nie *powiedziała*. Nawet wśród spójników trudno liczyć na porządek, skoro *and* tłumaczy się czasem jako *i*, czasem jako *a*. Nie mówiąc już o przedimkach *the* i *a*… I wreszcie czasowniki modalne w różnych formach: *was, had, have, would, will, were, been, are, could, do, should, has, must, can, did, may* czy *shall*, którym polszczyzna ma do przeciwstawienia już tylko *będzie, miał, miała* (z tym, że modalność nie jest jedyną funkcją tych słów) i *mogła*. Wspomniana już fleksyjność polszczyzny sprawia, że wiele cza-

sowników modalnych kryje się w polskich przyrostkach i dlatego znika z polskich list najczęstszych słów. Być może zresztą właśnie to zjawisko powoduje, że przekłady tekstów angielskich na mniej fleksyjny język francuski wykazują nieco silniejszy sygnał autora oryginału niż ich polskie odpowiedniki. I że sygnał autora był najsłabszy właśnie wśród polskich przekładów z francuskiego. I że te nieliczne przekłady, w których konsekwentnie dominował sygnał tłumacza, znalazły się w korpusie spolszczeń powieści francuskiej.

Chodzi tu przede wszystkim o chyba najciekawszy efekt widoczny na wykresach niniejszego rekonesansu: o wspólną gałąź Balzaka i Prousta, spolszczonych przez Tadeusza Boya Żeleńskiego. Ich odrębność trudno łączyć z czasem powstania, bo w tym korpusie znalazło się, oprócz wcześniejszych i późniejszych, wiele innych przekładów z tego samego okresu (m.in. i Balzaka). Trudno się oprzeć wrażeniu, że może nieprzypadkowo chodzi o zapewne najwybitniejszego (i najpracowitszego) polskiego tłumacza literatury francuskiej. A więc – popuszczając jeszcze bardziej wodze fantazji – warto zapytać, czy dobry tłumacz to czasem nie ten, który przestaje być niewidzialny, i który w dodatku pozostaje wierny własnemu profilowi stylometrycznemu, widocznemu także w jego własnej twórczości. Ale skoro tak, to – jeżeli wierzyć stylometrii – takich tłumaczy jest bardzo mało. I by być tłumaczem widzialnym (i to widzialnym nie dlatego, że popełnia się straszliwe błędy w tłumaczeniu), trzeba być… aż Boyem.

Bibliografia

Argamon S. 2008. *Interpreting Burrows's Delta. Geometric and Probabilistic Foundations*, „Literary and Linguistic Computing" 23 (2), 131–147.

Baayen R.H. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*, Cambridge: Cambridge University Press.

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

——— 2002a. *Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship*, „Literary and Linguistic Computing" 17, 267–287.

——— 2002b. *The Englishing of Juvenal. Computational Stylistics and Translated Texts*, „Style" 36, 677–699.

Chambers E.K. 1980. *The Problem of Chronology*, w: *William Shakespeare. A Study of Facts and Problems*, t. I, Oxford: Clarendon Press , 243–274.

Cooper Willis I. 1947. *The Authorship of* Wuthering Heights, „Trollopian" 2 (3), 157–168.

Craig H., Kenney A. 2009. *Shakespeare, Computers, and the Mystery of Authorship*, Cambridge: Cambridge University Press.

Dalen-Oskam K. van, Zundert, J. van. 2007. *Delta for Middle Dutch – Author and Copyist Distinction in Walewein*. „Literary and Linguistic Computing" 22, 345–362.

Eder M., Rybicki J. 2011. *Stylometry with R*, Digital Humanities Conference Abstracts, Stanford, 308–311.

Eder M., Rybicki J. 2013. *Do Birds of a Feather Really Flock Together, or How to Choose Test Samples for Authorship Attribution*, „Literary and Linguistic Computing" 28 (2) (w druku).

Eder M., Kestemont M., Rybicki J. (2012). *Computational Stylistics*, http://sites.google.com/site/computationalstylistics (dostęp: 1.04.2013).

Guedalla P. 1920. *Some Critics*, *Supers and Supermen*.

Hoover D.L. 2004. *Testing Burrows's Delta*. „Literary and Linguistic Computing" 19, 453–475.

——— 2007. *Corpus Stylistics, Stylometry, and the Styles of Henry James*, „Style" 41 (2), 174–203.

——— 2012a. *The Tutor's Story. A Case Study of Mixed Authorship*, „English Studies" 93 (3), 324–339.

——— 2012b. *The Rarer They Are, the More There Are, the Less They Matter*, Digital Humanities Conference Abstracts, Hamburg, 218–220.

Jockers M.L., Witten D.M., Criddle C.S. 2008. *Reassessing Authorship in the Book of Mormon Using Delta and Nearest Shrunken Centroid Classification*, „Literary and Linguistic Computing" 22, 465–491.

Lodge D. 1992. *Mały światek*, przeł. N. Billi, Poznań: Rebis.

Mosteller F., Wallace D.L. 1964. *Inference and Disputed Authorship. The Federalist*, Reading: Addison-Wesley.

Rybicki J. 2006. *Burrowing into Translation. Character Idiolects in Henryk Sienkiewicz's Trilogy and Its Two English Translations*, „Literary and Linguistic Computing" 21 (1): 91–103.

——— 2010. *Policzmy Trylogię Sienkiewicza*, „Annales Universitatis Paedagogicae Cracoviensis", 85–109.

——— 2011. *Ślady żony tłumacza. Alma Cardell Curtin i Jeremiah Curtin*, „Przekładaniec" 24, 90–110.

——— 2012. *The Great Mystery of the (Almost) Invisible Translator. Stylometry in Translation*, w: M. Oakes, M. Ji (red), *Quantitative Methods in Corpus-Based Translation Studies*, Amsterdam: John Benjamins, 231–248.

Rybicki J., Eder M. 2011. *Deeper Delta across Genres and Languages. Do We Really Need the Most Frequent Words?*, „Literary and Linguistic Computing" 26 (3), 315–321.

Rybicki J., Heydel M. 2013. *The Stylistics and Stylometry of Collaborative Translation. Woolf's Night and Day in Polish*, „Literary and Linguistic Computing" (w druku).

Smith P., Aldridge W. 2011. *Improving Authorship Attribution. Optimizing Burrows's Delta Method*, „Journal of Quantitative Linguistics" 18 (1), 63–88.

Strzetelski J. i in. 1987. *Historia literatury angielskiej. Tablice chronologiczne*, Warszawa: PWN.
Venuti L. 1995. *The Translator's Invisibility. A History of Translation*, Routledge, London–New York.
Vickers B. 2011. *Shakespeare and Authorship Studies in the Twenty-first Century*. „Shakespeare Quarterly" 62, 106–142.