Drugi rzut oka na stylometryczną mapę literatury polskiej

Jan Rybicki

Wprowadzenie

Pierwszy rzut oka na to, jak stylometryczne statystyki najczęstszych słów układają na płaszczyźnie mapę polskiego pisarstwa, prezentowany był kilka lat temu w tekście pod takim właśnie tytułem[1].W roku 2014 pięćset polskich tytułów wydawało się całkiem sporym wyborem tekstów. Dziś warto powrócić do tej samej myśli, bo w prezentowanych poniżej badaniach wykorzystano ponadpięciokrotnie większy zestaw, zawierający teksty polskie od *Kazań świętokrzyskich* po Szymiczkowej *Tajemnicę domu Helclów*, polskie powieści, poezję epicką i dramat oraz polskie przekłady z języka angielskiego (w tym liczne przekłady szekspirowskie), francuskiego, rosyjskiego, niemieckiego, czeskiego, włoskiego, portugalskiego, hiszpańskiego, węgierskiego i tureckiego.

Ktoś mógłby powiedzieć, że dotychczas nie czytano na dystans – od roku 2014 dorobiliśmy się też wreszcie spolszczenia tego okrzyczanego terminu, głównie za sprawą polskiego przekładu pierwszej książki twórcy tego pojęcia[2] – równie wielkiej kolekcji polskich tekstów. Zapewne nie – ale wcale nie jestem pewny, czy rzeczywiście chodzi tu o *distant reading*. Jak zauważa Matthew Jockers, jego były szef z Uniwersytetu Stanforda, Franco Moretti bada teksty „od zewnątrz" i „z daleka" przez statystyki wydawnicze, mapki wędrówek – rzeczywistych czy wirtualnych – autorów i postaci literackich i – jako dobry marksista – śledzi bardzo po darwinowsku ewolucję gatunków i form literackich, by stworzyć i promować swoistą genetykę literacką, nie unikając bezpośrednich nawiązań do badań DNA. Bardzo to oczywiście ciekawe, ale Jockers nie zgadza się rozszerzać terminu Morettiego na stylistykę komputerową, którą uprawiają wraz z nim – i jeszcze przed nim – tacy badacze, jak John Burrows, Hugh Craig, Karina van Dalen-Oskam, David Holmes, David Hoover Richard Forsyth czy Fotis Jannidis (a w Polsce Adam Pawłowski, Maciej Eder i piszący te słowa): „czytanie" wielu dzieł na raz za pomocą metod statystycznych, które do literaturoznawstwa tra-

[1] J. Rybicki, Pierwszy rzut oka na stylometryczną mapę literatury polskiej, „Teksty Drugie" 2014, nr 2, s. 106-128.

[2] F. Moretti, *Wykresy, mapy, drzewa. Abstrakcyjne modele na potrzeby literatury*, przeł. T. Bilczewski, A. Kowalcze-Pawlik, Kraków 2016.

fiły wprost z atrybucji autorskiej, i które są zorientowane na porównania bardziej „językowych" elementów tekstu, takich jak frekwencje słów, ich form podstawowych, części mowy. Dyscyplinę wiedzy, która dotychczas najchętniej występowała pod mianem stylometrii – ten termin zawdzięczamy Wincentemu Lutosławskiemu – Jockers nazwał sprytnie „makroanalizą"[3]. i nie wiadomo, czy tak już zostanie. W każdym razie warto pamiętać o tym rozróżnieniu, bo z niego wynikają inne zalety – i wady – obu tych skuzynowanych sposobów widzenia tekstów literackich.

Metoda

Główną wadą tego, co przynajmniej w tej chwili umawiamy się nazywać makroanalizą, jest odejście od tradycyjnie literaturoznawczego skupienia się na znaczeniu tekstu, jego zawartości czy przesłaniu. Jakże bo: oto badacz każe bezmyślnej maszynie pożreć tekst za tekstem, porąbać każdy tekst na pojedyncze słowa, a potem liczyć zmasakrowane, oderżnięte od wszelkiego kontekstu szczątki zdań, akapitów i rozdziałów, by ustalić listę słów najczęściej się powtarzających. Te zaś powtarzają się do znudzenia, bo każdy język naturalny – z delfinim włącznie – składa się przede wszystkim ze słów najkrótszych, najmniej „semantycznych" i najmniej określonych znaczeniowo (bo tak można streścić trzy rządzące tym rozkładem prawa Zipfa). W każdym więc języku naturalnym na stu pierwszych miejscach listy rangowej nie ma prawie żadnych słów o konkretnym znaczeniu (w badanym korpusie od biedy można znaleźć tam „oczy" i „domy"; toną jednak w powodzi różnych spójników, przyimków i zaimków) – ale za to stanowią one mniej więcej połowę każdego z tekstów. John Burrows już dawno pisał, że „badając [w sposób tradycyjny] dzieła literackie, zachowujemy się tak, jak gdyby jednej trzeciej, dwóch piątych czy nawet połowy materiału po prostu w nich nie było"[4], bo najczęściej nie zwracamy uwagi na tę językową tkankę łączną. A tymczasem – niestety– okazuje się, że właśnie statystyki tych „nieważnych" słów – a nie tych skrzydlatych, „znaczących" – najlepiej określają, „jak kto pisze". Oto bowiem – powracając na chwilę do taniej, horrorowej metaforyki – porąbawszy literackie arcydzieła na kawałki, stylometra wyciąga z nich tylko te najczęstsze, które najlepiej pasują mu do jego mrocznych celów, tworząc frankensteinowskiego potworka, który ma mu zastąpić istotę żywą: surową listę tych samych słów dla każdego zmasakrowanego tekstu.

Tu jednak – na szczęście – okazuje się, że choć listy te rządzą się tym samym rozkładem zipfowskim i są bardzo wszystkie do siebie podobne, to stanowią tak ogromną statystykę, że nieznaczne „na oko" różnice między nimi stają się bardzo znaczące, gdy zastosuje się do nich „szkiełko" analizy wielowymiarowej – przy czym wymiarów jest tyle, ile słów bierze się do analizy. Ile? Zwykle im więcej, tym lepiej, bo wiemy z całkiem podstawowej geometrii, że odległość między dwoma punktami na płaszczyźnie zwiększy się, jeżeli dodamy ich odległość w trzecim wymiarze, i że choć potem przestajemy „widzieć" – ludzkim okiem – dalsze wymiary, to odległość ta zawsze się zwiększy z każdym następnym wymiarem. Szkoda, że nie widzimy ich więcej – ale analiza wielowymiarowa jest po to, by zredukować wiele wymiarów do dwóch czy trzech, zwykle w wystarczająco dobrym przybliżeniu, by „oddać" skalę odległości – czyli po prostu różnic – w stosowaniu tych najczęstszych słów przez różnych indywidualnych autorów, ich grupy artystyczne, pokoleniowe czy genderowe, epoki, gatunki i rodzaje literackie. I tak

[3] M. Jockers, *Macroanalysis. Digital Methods and Literary History*, Champaign 2013.

[4] J. Burrows, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, Oxford 1987, s. 1.

właśnie powstają „wykresy, mapy, drzewa", ale inne niż te u Morettiego, bo oparte wyłącznie na surowym materiale językowym. Ale to wystarczy, bo właśnie one przybierają często kształty, które zwykle zaskakują tradycyjnego literaturoznawcę w ich podobieństwie do tradycyjnych odczytań. Warto dodać, że choć obowiązuje zasada „im więcej, tym lepiej", w praktyce statystyka wysyca się – a wyniki stają się bardzo stabilne – na poziomie 1000-2000 słów[5].

Sęk w tym, że choć istnienie takiego autorskiego czy chronologicznego „odcisku palca" znalazło liczne empiryczne potwierdzenia, sam mechanizm powstawania tych podobieństw i różnic między tekstami nie jest dostatecznie wyjaśniony przez językoznawstwo, i tylko jego kognitywna gałąź zerka czasem z ciekawością na makroanalityczne pomysły[6]. Oczywiście fakt, że każdy piszący używa – na pewno częściowo nieświadomie – tych wspólnych najczęstszych słów w jakichś własnych, indywidualnych proporcjach, nie kłóci się zbyt zajadle z intuicją; a już na pewno pisarze piszący w tych samych epokach stosują się do kształtu języka na wspólnym etapie rozwoju. Gorzej (bo jakoś trudno to zrozumieć), że stylometryczny sygnał autora potrafi przetrwać nawet traumę przekładu na inny język. Choć badanie oryginałów w języku wyjściowym i przekładów w języku docelowym dokonuje się przecież na dwóch rozłącznych listach frekwencyjnych, w których na próżno szukać dokładnych odpowiedniości między np. przyimkami w jednym i w drugim języku, wykresy czy mapy sporządzone na ich podstawie najchętniej grupują teksty właśnie według autora oryginału, a nie tłumacza[7]. Różnych tłumaczy trochę łatwiej rozpoznać, gdy przekładają tego samego autora lub nawet ten sam tekst[8].

A w dodatku żadne inne klasyfikatory nie sprawdzają się równie dobrze jak te nieszczęsne listy częstych słów: ani słowa kluczowe, ani n-gramy (czyli sekwencje) sąsiadujących słów, ani nawet n-gramy słów tych wartości (tagów) gramatycznych nie dają zwykle równie czystego obrazu autorstwa czy chronologii[9]. Nawet jeżeli czasem dają podobne wyniki, to są znacznie bardziej kłopotliwe w przetwarzaniu komputerowym. Prosta lemmatyzacja (sprowadzanie wszystkich wyrazów do ich form podstawowych) też nie polepsza znacząco wyników (nic dziwnego zresztą, skoro to stylistycznie bardzo znacząca decyzja, gdy autor wybiera np. narrację w czasie teraźniejszym, a nie w przeszłym). Stylometria wprawdzie nie ustaje w próbach przerwania tej dominacji leksyki – między innymi po to, by uczynić swoje badania bardziej „strawnymi" dla tradycyjnego literaturoznawstwa z jednej, a językoznawstwa z drugiej strony – ale z nikłymi na razie rezultatami. Jeden tylko sygnał genderowy przejawia się w bardziej „znaczących" słowach, gdy poszukuje się go w narodowych literaturach XVIII i XIX wieku[10].

[5] J. Rybicki, M. Eder, *Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?*, „Literary and Linguistic Computing" 2011, nr 26 (3), s. 315-332; M. Eder, *Does size matter? Authorship attribution, small samples, big problem*, „Literary and Linguistic Computing" 2015, nr 30 (2), s. 167-182.

[6] Warto wspomnieć o pracy doktorskiej australijskiej badaczki Louisy Connors, *Computational Stylistics, Cognitive Grammar, and the Tragedy of Mariam: Combining Formal and Contextual Approaches in a Computational Study of Early Modern Tragedy*, Newcastle 2013.

[7] J. Rybicki, *The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation*, [w:] *Quantitative Methods in Corpus-Based Translation Studies*, red. M. Oakley, M. Ji, Amsterdam 2012, s. 231-248.

[8] J. Rybicki, 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.

[9] R. Górski, M. Eder, J. Rybicki. *Stylistic fingerprints, POS tags and inflected languages: a case study in Polish*, [w:] *Qualico 2014: Book of Abstracts*, Olomouc 2014, s. 51-53.

[10] J. Rybicki, *Vive la différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies*, „Digital Scholarship in the Humanities" 2016, nr 31 (4), s. 746-761.

Skoro nie ma więc w tej chwili nic „lepszego" – czy przynajmniej nic bardziej „strawnego" – czas przedstawić pokrótce, skąd biorą się ukazane poniżej wizualizacje. Dokładniejszy opis całej procedury znaleźć można w polskim tekście Macieja Edera[11] i w tego autora znacznie bardziej „technicznym" artykule opublikowanym w „Digital Scholarship in the Humanities"[12]. Z kolei opis strony programistycznej metody zamieścił prestiżowy „R Journal"[13]: większość etapów obliczeniowych wykonywana jest za pomocą opisanego tam pakietu stylometrycznego „stylo", napisanego dla środowiska oprogramowania statystycznego R[14]. Pakiet pobiera wersje elektroniczne wszystkich tekstów, dzieli je na słowa, zlicza ich częstości w całym korpusie, wybiera określoną przez użytkownika liczbę najczęstszych spośród nich i uzyskawszy w ten sposób ciągi liczb dla każdego z tekstów, porównuje te ciągi dla każdej pary tekstów. Porównanie to zasadza się na wyznaczeniu miary odległości – różnicy – między tekstami. Spośród licznych miar stosowanych w stylometrii w niniejszym badaniu zastosowałem tę, która wykazuje największą skuteczność w wykrywaniu sygnału autorskiego: wariant Delty Burrowsa[15] wykorzystujący kosinus kąta między wektorami częstości słów dla każdej pary tekstów[16]. „Delta kosinusowa" ($\Delta\angle$) dla dwóch tekstów T i T_1 mierzy więc kąt α (im większy, tym większa różnica między tymi tekstami):

$$\Delta\angle(T, T_1) = \alpha,$$

wyliczany na podstawie podobieństwa kosinusowego tzw. standaryzacji Z dwóch wektorów $x = z(T)$ i $y = z(T_1)$:

$$\cos \alpha = \frac{\sum_{i=1}^{n_s} x_i y_i}{\sqrt{(\sum_{i=1}^{n_s} x_y^z)}\sqrt{(\sum_{i=1}^{n_s} y_i^z)}},$$

gdzie n_s to liczba badanych słów, a z(T) to wartość standaryzacji Z częstości słów w tekście T, liczona zwykłym wzorem:

$$z\,(T) = \frac{f_s(T) - \mu_s}{\sigma_s},$$

gdzie z kolei $fs(T)$ to częstość bezwzględna słowa s w tekście T, μs to średnia częstość słowa s w zbiorze tekstów, do którego należy tekst T, a σs to odchylenie standardowe częstości słowa

[11]M. Eder, *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii*, „Teksty Drugie" 2014, nr 2, s. 90-105. Studium to stanowiło świadomy teoretyczno-praktyczny tandem z moim cytowanym powyżej tekstem, zamieszczonym w tym samym czasopiśmie.

[12]M. Eder, *Visualization in Stylometry: Cluster Analysis Using Networks*, „Digital Scholarship in the Humanities" 2017, nr 32 (1), s. 50-64.

[13]M. Eder, J. Rybicki, M. Kestemont, *Stylometry with R: A Package for Computational Text Analysis*, „R Journal" 2016, nr 8 (1), s. 107-121.

[14]R Core Team. *R: A language and environment for statistical computing*, <http://www.R-project.org/> 2014 [dostęp: 14.07.2017].

[15]J. Burrows, *Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship*, „Literary and Linguistic Computing" 2002, nr 17, s. 267-287.

[16]P.W.H. Smith, W. Aldridge, *Improving Authorship Attribution: Optimizing Burrows' Delta Method*, „Journal of Quantitative Linguistics" 2011, nr 18 (1), s. 63-88.

s w tym samym zbiorze tekstów[17]. Uzyskuje się w ten sposób miarę odległości Δ∠ dla każdej pary tekstów i w efekcie dla całego zbioru tekstów powstaje macierz odległości w całym korpusie. Już na tej podstawie można wnioskować o tym, które teksty są do siebie podobne, ale znacznie czytelniejszy obraz tych relacji dla bardzo dobrej próbki – nie bójmy się tego terminu w przypadku niniejszych badań – polisystemu literatury po polsku dadzą nam dwuwymiarowe wizualizacje, sporządzane za pomocą wybranych metod statystycznych.

Macierz odległości może być – i tak czyniłem w przypadku opisywanego studium – zbadana analizą skupień, która łączy ze sobą najbardziej podobne teksty w ramach całej kolekcji. I tak na przykład *Ogniem i mieczem* zostaje uznane za najbliższego sąsiada *Potopu*; następnym najbliższym sąsiadem obu zostaje z kolei *Pan Wołodyjowski*. Można się spodziewać, że w następnej kolejności skupisko trzech części tej samej trylogii połączy się z *Krzyżakami*, potem z *Quo vadis* i jeszcze później z *W pustyni i w puszczy*; należy się spodziewać (i tak jest w rzeczywistości), że takie większe skupisko sienkiewiczowskiej literatury przygodowej wtedy dopiero spotka się z *Bez dogmatu* i *Rodziną Połanieckich*, pełny zaś Sienkiewicz połączy się z podobnie budowanym pełnym Prusem w kolejnym stadium łączenia tekstów na podstawie podobieństwa stylometrycznego. W ten sposób wielowymiarowa przestrzeń stworzona przez teksty całego korpusu i wszystkie użyte w analizie słowa zostaje zredukowana do czegoś, co można przedstawić na płaszczyźnie.

Wśród metod takiego przedstawienia wielką karierę zrobiła ostatnio tak zwana analiza sieciowa, która rozkłada w dwóch lub trzech wymiarach punkty danych (w tym przypadku poszczególne teksty) w zależności od stopnia podobieństwa: im większe podobieństwo, tym dwa punkty dzieli mniejsza odległość, a łączy grubsza linia. Oczywiście dla bardzo wielu tekstów potrzeba do tego dużo matematyki. Za badacza całość pracy wykonuje program Gephi[18] za pomocą grawitacyjnego algorytmu Force Atlas 2. Według jego twórców algorytm „symuluje fizyczny układ w celu umieszczenia sieci w płaszczyźnie. Węzły [sieci, czyli punkty poszczególnych tekstów] odpychają się jak tak samo naładowane cząstki, ale krawędzie [sieci, czyli miary podobieństwa między węzłami] przyciągają je do siebie jak sprężyny, by w końcu osiągnąć stan stałej równowagi”[19]. W naszym przypadku Gephi pobiera z wyników „stylo” liczby ukazujące, jak często dana para tekstów znajduje się w bliższym lub dalszym sąsiedztwie; to częstość tych „zetknięć” stanowi o sile podobieństwa między dwoma tekstami, o sile owej sprężyny, nie pozwalającej dwóm tekstom oddalić się od siebie zbyt daleko. Jak już wspomniałem powyżej, prędzej czy później (w zależności od rozmiarów sieci i mocy procesora) układ osiąga stan równowagi. Sieć powstaje. „Mapa” (w tym przypadku) literatury polskiej jest gotowa. Można dopatrywać się w niej odrębnych skupisk na dwa sposoby: albo na podstawie tradycyjnej wiedzy historyczno-literackiej przypisuje się poszczególne teksty do autorów, epok, gatunków,

[17]S. Evert, T. Proisl, F. Jannidis, I. Reger, S. Pielström, C. Schöch, T. Vitt, *Understanding and explaining Delta measures for authorship attribution*, „Digital Scholarship Humanities” 2017 <https://academic.oup.com/dsh/article-abstract/doi/10.1093/llc/fqx023/3865676/Understanding-and-explaining-Delta-measures-for> [dostęp: 14.07.2017].

[18]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.

[19]M. Jacomy, T. Venturini, S. Heymann, M. Bastian, *ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software*, „PLoS ONE” 2014, nr 9(6), e98679, doi:10.1371/journal.pone.0098679.

przedziałów czasowych, albo podziału dokonuje się matematycznie, stosując funkcję modularności. Dla sieci „ważonych" – czyli takich, jak te sporządzane w tym studium, a więc takich, w których połączenia między poszczególnymi węzłami mają różne „wagi" –modularność sieci oblicza się wzorem:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

gdzie *Aij* to właśnie waga („siła") połączeń (podobieństw) między punktami (tekstami) *i* oraz *j*; *ki = ΣjAij* to suma wag wszystkich połączeń zbiegających się w węźle *i*; *ci* to skupisko, do którego zostaje przydzielony węzeł *i*; wreszcie funkcja *δ(u, v)* przybiera wartość 1, gdy *u = v* i wartość 0 gdy *u ≠ v*, zaś *m = ½ΣijAij*[20]. Można powiedzieć, że powyższy wzór jest tym, czym komputer zastępuje ludzką wiedzę o autorach, epokach, gatunkach literackich.

Materiał

Zanim przejdę do wyników, warto rozszerzyć nieco opis korpusu. Największą grupę (1319 tytułów) stanowią polskie oryginały powieści, poematów epickich i – szczególnie wśród najstarszych tekstów – kazania, psałterze i żywoty świętych. Oczywiście panuje tu silna dysproporcja na korzyść wszelkich gatunków prozy powieściowej, na co składają się dwa związane zresztą ze sobą czynniki: po pierwsze, powieści jest po prostu najwięcej, po drugie, to właśnie one są najbardziej dostępne w wersjach elektronicznych. Podobnie reprezentatywna jest druga dysproporcja – chronologiczna. Wiek XIV reprezentowany jest przez jeden tekst; wiek XV i XVI wprowadziły do korpusu odpowiednio po dziesięć i dziewięć tekstów. Następny wiek nie na darmo nazywa się „wiekiem rękopisów" – liczba dostępnych tekstów maleje do ośmiu; ale jeszcze gorzej jest w przypadku wieku XVIII: to zapewne dominacja gatunków nieepickich sprawia, że mimo wysiłków pewnego biskupa warmińskiego jest tu tylko tekstów pięć. Eksplozja twórczości powieściowej – i jej elektronicznej dostępności – ujawnia się w wieku XIX, przynosząc 426 tytuły, i nasila się jeszcze w wieku XX (631 tekstów). Na tym tle całkiem dzielnie poczyna sobie młodziutkie nowe tysiąclecie, bo jego pierwsze kilkanaście lat może się pochwalić 229 tytułami. Trudno się dziwić – to przecież XXI wiek wprowadza literaturę w medium elektroniczne często bez pośrednictwa druku na papierze.

Tyle o prozie i epice polskiej. Osobno liczę polski dramat od Kochanowskiego po Mrożka. Są to 63 teksty: poza *Odprawą posłów greckich* oczywiście Fredro, cała trójka wieszczów, „ciężkie Norwidy" i dużo Wyspiańskiego; oczywiście Zapolska, Przybyszewski i Witkacy; oprócz pojedynczych tekstów autorów spore grupy tworzą Gombrowicz i Mrożek. Łącznie wszystkich tekstów rdzennie polskich jest 1382. Warto dodać, że przeczytanie ich wszystkich – nawet w zawrotnym tempie jednego w dwa dni – zabrałoby jednej osobie ponad siedem i pół roku.

Ale oto nadchodzą przekłady z innych języków. Skoro od międzywojnia najwięcej tłumaczy się w Polsce z angielskiego[21], przekładów z języka Byrona jest w prezentowanej kolekcji 408 – ale

[20] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, *Fast Unfolding of Communities in Large Networks*, „Journal of Statistical Mechanics: Theory and Experiment" 2008, nr 10, s. 1000.

[21] Por. W. Krajewska, *Recepcja literatury angielskiej w Polsce w okresie modernizmu (1887–1918). Informacje. Sądy. Przekłady*, Wrocław–Warszawa–Kraków–Gdańsk 1972.

to bez Szekspira, któremu wypada policzyć osobno aż 135 przekładów; razem jest ich 543. „Francuzów" jest znacznie mniej – 242, „Rosjan" – 103 i dokładnie tyleż „Niemców"; Czesi, Hiszpanie, Węgrzy, Włosi, Skandynawowie i Turcy wprowadzili łącznie 175 tekstów. Literatury obcej (1161) mamy więc stosunkowo niewiele mniej niż polskiej; łącznie zaś jest tytułów 2548. Rozmiar całego korpusu to 170 692 206 słów.

Skąd one wszystkie wzięły się w formie elektronicznej? Dokładnej statystyki niestety nie zbierałem. Znaczna część prac – tych w domenie publicznej – pochodzi z różnych darmowych kolekcji, szlachetnych i pożytecznych przedsięwzięć w rodzaju *Wolne lektury*[22], *Biblioteka literatury polskiej w Internecie*[23] czy *Staropolska*[24]. Te trzy repozytoria były najbardziej przydatne; najstarsze polskie teksty pochodziły z małej, ale bezcennej elektronicznej „Biblioteki zabytków polskiego piśmiennictwa średniowiecznego" Instytutu Języka Polskiego PAN w Krakowie[25]. Teksty nowsze przyszły często wprost z księgarni internetowych w formie e-booków – to oczywiście przyspiesza pozyskiwanie tekstów i w dodatku obniża koszty, bo książki elektroniczne są często (nieznacznie) tańsze od papierowych. Duża część jednak musiała zostać przeniesiona w medium elektroniczne przez skanowanie i OCR[26]. Niedawne wyposażenie Instytutu Filologii Angielskiej Uniwersytetu Jagiellońskiego w skanery z podajnikiem umożliwiło niemal bezinterwencyjną cyfryzację książek – pod warunkiem, że każdy wolumin został najpierw rozdzielony na osobne kartki[27].

W tym miejscu krótka dygresja o stanie dostępności literatury w języku polskim – oryginalnej i spolszczonej – w formie elektronicznej. Skoro do niniejszego studium udało się pozyskać ponad dwa tysiące tekstów, ktoś mógłby pomyśleć, że nasza literatura opanowała już medium cyfrowe. Z punktu widzenia „zwykłego" czytelnika jest to nawet dość prawdziwe: przeczytać książkę w Internecie lub z Internetu jest rzeczywiście całkiem łatwo. Gorzej z pozyskaniem tekstu do analizy ilościowej, bo niemal każde repozytorium stosuje inny format, inny interfejs użytkownika i – co zrozumiałe – na ogół dość skutecznie broni się przed udostępnianiem całości czy większych partii swoich zasobów. Tradycyjnemu czytelnikowi nie przeszkadza, że musi czytać pierwszy polski przekład Hamleta (Wojciech Bogusławski, 1797, na podstawie niemieckiej adaptacji Friedricha Ludwiga Schrödera) w bardzo niewygodnym do obróbki formacie DjVu. Wręcz przeciwnie, *le plaisir du texte*, w którym tytułowy bohater na szczęście przeżywa i któremu towarzyszy nie Horacy, a Gustaw, jest tym większa, że na ekranie komputera pojawia się piękny starodruk z 1823 roku. Tradycyjny czytelnik jakoś

[22] <http://wolnelektury.pl> [dostęp: 14.07.2017].

[23] <http://literat.ug.edu.pl/> [dostęp: 14.07.2017].

[24] <http://www.staropolska.pl/> [dostęp: 14.07.2017].

[25] *Biblioteka zabytków polskiego piśmiennictwa średniowiecznego*, red. W. Twardzik, Kraków 2006.

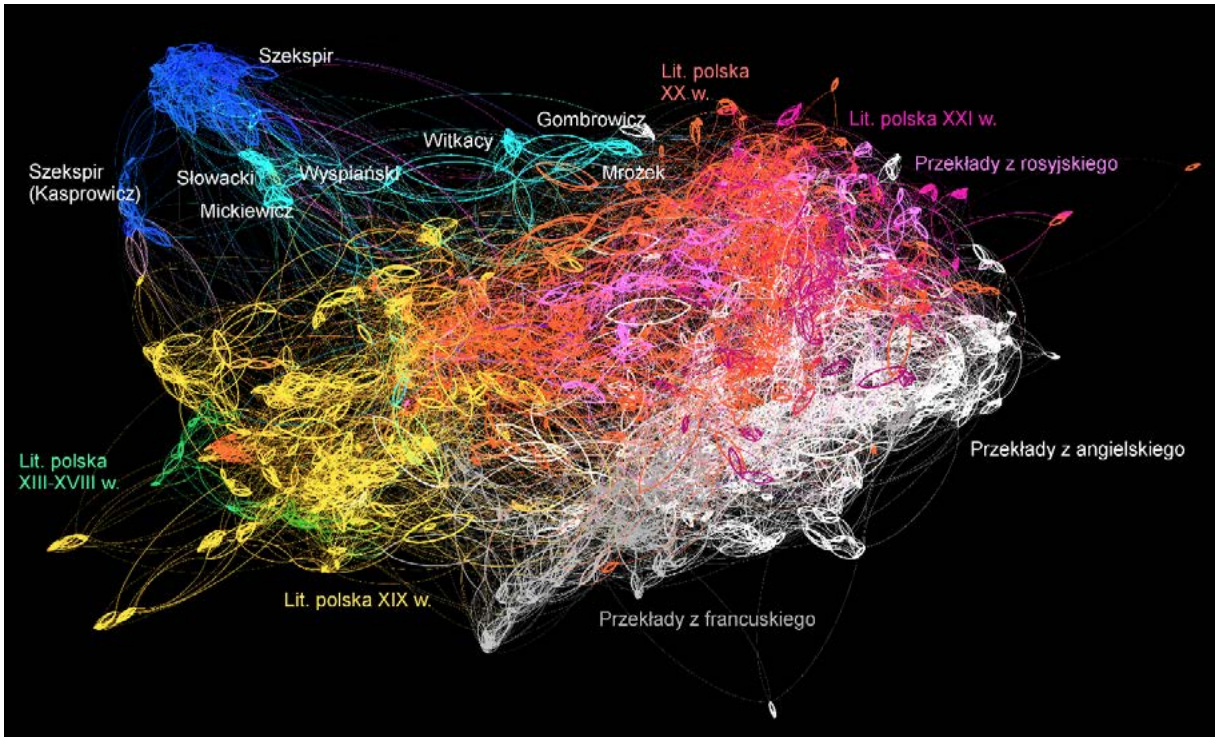
[26] Nie sposób tu nie wspomnieć o heroicznym trudzie moich dwóch magistrantek, Anny Hołubiczko i Marty Kamudy, które zebrały tak imponujący korpus polskich przekładów Szekspira, zawzięcie skanując wersje papierowe lub z mniszą precyzją poprawiając trudne, bo starodrukowe skany Biblioteki Polona (<https://polona.pl/> [dostęp: 14.07.2017]). Efektem tych starań – poza korpusem szekspirowskim i znacznym wkładem w podkorpus dramatu polskiego – są dwie ciekawe prace magisterskie: A. Hołubiczko, „«Porównania śmierdzą»: porównanie równoległych tekstów polskich przekładów Szekspira", Kraków 2017; M. Kamuda, „Stylometric Analysis of the Polish Translations of Shakespeare", Kraków 2017.

[27] I tu wielkie podziękowania należą się Pracowni Introligatorskiej Volumin przy ulicy Św. Gertrudy 5 w Krakowie, która – z żalem, bo z żalem, ale jednak – darmowo rozcinała każdy wolumin w zamian za niejasną obietnicę złożenia go kiedyś z powrotem.

da sobie radę nawet w sytuacji, gdy niektóre z rzekomo scyfryzowanych tekstów w polskich zasobach to w rzeczywistości obrazkowe PDF-y, które dopiero wymagają rozpoznania w nich tekstu. Mała pociecha, że tak dzieje się nie tylko w Polsce, i to mimo istnienia żelaznych – wydawałoby się – zasad rządzących kodowaniem tekstu, wypracowywanych przez konsorcjum Text Encoding Initiative. Humanistyka cyfrowa na całym świecie, której najbardziej rozpoznawana „działka" to właśnie tworzenie cyfrowych repozytoriów wszelkich artefaktów kulturowych, potrafi tworzyć piękne i cenne wydania cyfrowe dla użytkownika „z zewnątrz", ale wciąż jakby zapomina o ważnych klientach z własnego środowiska – tych właśnie, którzy zajmują się ilościową analizą danych kulturowych. A przecież według Willarda McCarty'ego, jednego z najwybitniejszych autorytetów cyfrowego zwrotu w badaniach humanistycznych, to właśnie stylistyka komputerowa najbardziej przyczynia się do tego zwrotu i w największym stopniu wskazuje nowe drogi[28].

Wyniki

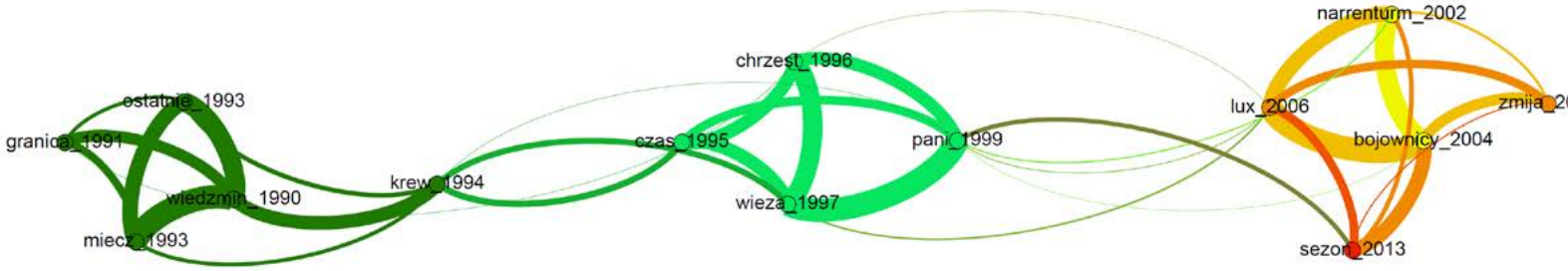
Jak więc wygląda – w takim makroanalitycznym ujęciu – reprezentacyjny wycinek literatury w języku polskim? Tak, jak na wykresie 1.



Wykres 1. Analiza sieciowa 2548 tekstów na podstawie częstości 2000 najczęściej występujących słów w całym korpusie

[28]W. McCarty, *Getting There from Here. Remembering the Future of Digital Humanities: Roberto Busa Award Lecture 2013*, „Literary and Linguistic Computing" 2014, nr 29 (3), s. 197.

Podobnie jak w cytowanej powyżej makroanalizie znacznie mniejszego korpusu literatury polskiej[29], sieć wykazuje silne uporządkowanie chronologiczne w partiach polskich oryginałów – nawet jeżeli na odległych orbitach wykresów pojawiają się mniej zdyscyplinowane satelity. Teksty wczesne, oznaczone kolorem zielonym, grupują się na ogół w lewym dolnym rogu wykresu. Literatura XIX wieku (kolor żółty) jest przesunięta w prawo i w górę, po czym powoli przechodzi w wiek XX (czerwień). Ciemnofioletowe skupiska w prawym górnym rogu wykresu to pisarstwa wieku XXI. Najważniejszym spostrzeżeniem dla tego elementu wizualizacji jest zresztą nie tyle istnienie skupień chronologicznych, co ich ewolucyjna progresja w jednym kierunku. W literaturze przedmiotu od dawna zwracano uwagę na takie stopniowe, obdarzone takim samym zwrotem „drobne kroczki ku nieskończoności" (*tiptoeing towards the Infinite*) i dość skutecznie przekonywano, że jest to wynik nie tylko zmian językowych, że jest to również efekt ewolucji stylometrii – jeśli nie stylistyki – literackiej[30]. Najlepszym chyba argumentem na potwierdzenie takiego odczytania wizualizacji jest występowanie ewolucyjnych, ukierunkowanych trendów w obrębie twórczości pojedynczego autora, gdzie istnienie znaczących zmian w samej tylko polszczyźnie staje się trudniejsze do obronienia. Jako dobry przykład może posłużyć analiza sieciowa twórczości Andrzeja Sapkowskiego (wykres 2), którą częstości najczęstszych słów dzielą na trzy wyraźne, kilkuletnie okresy.



Wykres 2. Periodyzacja twórczości Andrzeja Sapkowskiego na podstawie częstości najczęstszych słów: pierwsza połowa lat dziewięćdziesiątych XX wieku (ciemna zieleń); druga połowa lat dziewięćdziesiątych (jasna zieleń); wiek XXI (żółć)

Powróćmy jednak do wykresu 1, ponieważ dzieją się tam jeszcze inne ciekawe rzeczy. Oto bowiem gdy literatura polska (a raczej jej epicko-powieściowy *mainstream*) płynie z lewa na prawo, doczepia się do niej od dołu spora szaro-biała masa. Linie szare łączą ze sobą polskie przekłady literatury francuskiej, linie białe to przekłady z angielskiego. Jeżeli chronologiczny sygnał jest wspólną cechą całego wykresu, trudno nie powiązać pojawienia się przekładów z francuskiego „wcześniej", bo bardziej na lewo z wcześniejszym oddziaływaniem na twórczość literacką w Polsce pisarzy francuskich. Anglicy i Amerykanie lądują później – bardziej na prawo – ale za to jakby skuteczniej – choć i tak nie do końca – udaje im się zinfiltrować

[29] J. Rybicki, *Pierwszy rzut oka*...

[30] J. Burrows, *Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative*, [w:] *Research in Humanities Computing 4*, red. S. Hockey, N. Ide, Oxford 1996, s. 1-33.

obszary rdzennej literatury polskiej. Nie bez znaczenia może być fakt, że przekłady z francuskiego w analizowanym korpusie są z powyższych względów nieco wcześniejsze niż te z języka angielskiego.

To jednak nie jedyne ciekawostki przekładoznawcze tej wizualizacji. Wśród szarego morza spolszczeń literatury francuskiej można zauważyć kilka białych wysepek: to przekłady twórczości Waltera Scotta. Zresztą na szaro-białym pograniczu znajdują się też wczesne przekłady Dickensa. Chodzi w tym przypadku o dwóch powieściopisarzy angielskich, którzy najwcześniej zdobyli uznanie nad Wisłą. Można się domyślać na tej podstawie, że układ przekładów powieściowych również zawdzięczamy oddziaływaniu czynnika chronologicznego.

Nie jest to jednak jedyny czynnik wpływający na skład skupisk w tym wykresie sieciowym. Pierwszy polski przekład arcydzieła Charlotte Brontë, *Janina* – bo tak w roku 1880 zdecydowała się spolszczyć tytuł *Jane Eyre* tłumaczka Emilia Dobrzańska – woli trzymać się szarych Francuzów niż białych Anglików. Nic dziwnego: ta polska wersja jest nie tylko skrócona, ale w dodatku stworzona na podstawie francuskiego przekładu, o czym – w *close reading* – świadczą liczne kalki z francuskiego[31]. W podobnej sytuacji znajduje się kilka innych starych przekładów z angielskiego, co do których żywić można podobne podejrzenia. W ten sposób „czytanie na dystans" może wskazywać ciekawe tematy do czytania „z bliska".

I wreszcie: w tej samej francuskiej szarzyźnie czerwienią się – w skupisku Stendhali, Balzaków i Proustów przełożonych przez Tadeusza Boya-Żeleńskiego tegoż *Znaszli ten kraj* i *Marysieńka Sobieska*. W ten sposób polski lekarz trafia do grona tłumaczy, których stylometryczny odcisk palca nie zmienia się niezależnie od tego, czy piszą własne, czy tłumaczą cudze. Nie pierwszy raz badania ilościowe wskazują na tę właśnie cechę pisarstwa Boya[32] i nie on jeden ją posiada. Ale są też autorzy, którzy tłumaczą zupełnie inaczej, niż piszą: tłumacząc Juwenalisa z łaciny na angielszczyznę Samuel Johnson, jak Boy, „trzyma się własnej nuty", podczas gdy John Dryden „znajduje dla swego przekładu całkiem nowy ton"[33].

O ile dwie wielkie literatury niesłowiańskie zaledwie ocierają się o główny korpus polskich oryginałów, o tyle jasnofioletowe smugi przekładów z rosyjskiego przenikają w sam środek czerwonego obszaru literatury polskiego wieku XX. Z kolei rosyjska fantastyka naukowa miesza się z ciemnym fioletem najnowszej literatury polskiej, wśród której przecież nie brakuje przedstawicieli tego gatunku. Sygnał gatunkowy objawia się więc w sposób bardzo dla siebie typowy[34]. Ale inne zachowanie przekładów z innego języka słowiańskiego sugeruje istnienie interesujących wahań w nasileniu *translationese*, które zdaje się rosnąć wraz z różnicami między językiem wyjściowym a docelowym – tym bardziej że nieliczne niestety w badanym korpusie (i dlatego na wykresie nie zaznaczone) przekłady z czeskiego zachowują się podobnie do tłumaczeń z ruszczyzny.

[31] D. Hadyna, „A controversial translation justified by the context: Janina, the first Polish version of Charlotte Brontë's Jane Eyre" (praca magisterska), Kraków 2013.

[32] J. Rybicki, *Stylometric Translator Attribution: Do Translators Leave Lexical Traces?*, [w:] *The Translator and the Computer*, red. T. Piotrowski, Ł. Grabowski, Wrocław 2013, s. 193-204.

[33] J. Burrows, *The Englishing of Juvenal: Computational Stylistics and Translated Texts*, „Style" 2002, nr 36, s. 677-699.

[34] Por. C. Schoech, *Fine-tuning our stylometric tools: Investigating authorship, genre, and form in French classical theater*, [w:] *Digital Humanities 2013 Conference Abstracts*, red. K. Walter, K. Price, Lincoln 2013, s. 383-386.

Jednak najciekawszym efektem związanym z przekładem jest olbrzymi dystans między – ukazanymi na ciemnoniebiesko – przekładami z Szekspira a położoną na przeciwległym krańcu sieci białą plamą reszty spolszczonej literatury angielskiej. Oczywiście jednym z powodów takiego rozgraniczenia są różnice rodzajowe, bo „białe" teksty to wyłącznie proza powieściowa. Nie zmienia to jednak faktu, że polski Szekspir rządzi się własnymi prawami. Choć badany korpus zawiera efekty pracy aż 19 różnych tłumaczy angielskiego barda, ma ten polski Szekspir swój własny stylometryczny profil. Od tej reguły wyłamują się tylko – i tylko w pewnym stopniu – przekłady Kasprowicza i niektórych innych tłumaczy z *fin de siècle*. A że niedaleko – całkiem naturalnie – przebiega jasnoniebieski szlak polskiego dramatu (którego chronologia ma ten sam zwrot, co reszta literatury polskiej: od lewej do prawej), sfera szekspirowska najsilniej przyciąga te jego elementy, które wpływ Szekspira mają niejako na sztandarze i w manifeście: romantyczny dramat Mickiewicza i Słowackiego, a tuż obok neoromantyczny teatr autora *The Tragicall Historie of Hamlet Prince of Denmark. Według tekstu polskiego Józefa Paszkowskiego, świeżo przeczytana i przemyślana przez St. Wyspiańskiego*[35].

Wszystkie te obserwacje łączy jedna wspólna zasada: o ile maszyna zajmuje się liczeniem i samą grafiką, podział punktów wykresu i ich klasyfikacja jest wciąż jak najbardziej „ludzka", humanistyczna. Człowiek-interpretator wie przecież, który punkt oznacza który tekst (nawet jeżeli może mieć kłopoty z odnalezieniem go w gąszczu wielkich sieci), i sam decyduje o odrębnym zabarwieniu Szekspira, literatury polskiej XX wieku, przekładów z angielskiego itd. Obraz, który powstaje w ten sposób, musi z natury rzeczy być silnie zależny od tradycyjnej historii literatury, która też jest głównym punktem odniesienia w ocenie wizualizacji komputerowej. Wizualizacja może wskazywać interesujące nieciągłości czy nieoczekiwane z punktu widzenia tradycyjnej nauki o literaturze połączenia – albo właśnie ich brak. W takiej interpretacji człowiek narzuca sam sobie choćby samą liczbę i naturę klas, na które dzieli badany materiał: na autorów, epoki, gatunki, języki wyjściowe.

Maszyna może jednak wyręczyć człowieka w jednej z tych czynności. Maszynę można poprosić o to, by sama podzieliła badane teksty na pożądaną liczbę grup. Człowiek nadal co prawda decyduje, ile będzie tych grup, ale podziały mogą, choć nie muszą, przebiegać zupełnie inaczej niż te wynikające z ludzkiej wiedzy o analizowanych tekstach. W tym celu warto wykorzystać wspomnianą powyżej funkcję modularności w pakiecie Gephi.

Zobaczmy więc, co stanie się, jeżeli komputer spróbuje nam wskazać – oczywiście na podstawie najmniejszych różnic w użyciu najczęstszych słów – jak podzielą się badane dzieła, jeżeli dopuścimy istnienie dwóch – i więcej – głównych grup. Wykres 3 prezentuje zestawienie takich wizualizacji dla 2, 3, 4 i… 70 grup.

[35]Kraków 1905.

Wykres 3. Analiza sieciowa korpusu z modularnym podziałem na (od lewej i od góry) 2, 3, 4 i 70 skupisk

Mając możliwość podziału tekstów tylko na dwie grupy, algorytm modularności dzieli badany korpus na wspólne, wielkie skupisko prozy w oryginale i w przekładzie (zieleń) oraz na dramat (fiolet) polski i szekspirowski. Do tej drugiej grupy dołącza część wczesnej powieści polskiej z połowy XIX wieku (Duchińska, Goszczyński, Niewiarowski, Michał Jezierski). Trójpodział wprowadza grupę prozy wcześniejszej (zieleń; mniej więcej do połowy XX wieku) i prozy późniejszej wraz z większością przekładów (fiolet). Dramat polski i szekspirowski (bez Kasprowicza i jego sąsiadów z tej samej epoki) to osobne, żółte skupisko. Dalsze zwiększanie liczby grup prowadzi powoli do wyłonienia się wielu grup autorskich. Dopiero jednak przy 70 grupach udaje się oddzielić polski dramat romantyczny i neoromantyczny od Szekspira i jest to interesująca miara podobieństwa językowego między tymi silnie literacko spowinowaconymi kategoriami.

Wnioski

Redaktor niniejszego zbioru nazwał wykres 1 „nieznanym Pollockiem”[36] – i rzeczywiście nie da się ukryć, że opis i komentarz do wizualizacji analizy sieciowej ponad dwóch tysięcy tekstów zaczyna mieć ekfrastyczne konotacje. Można nawet powiedzieć, że mamy do czynienia z interesującą i dotąd w badaniach kultury rzadko spotykaną transformacją: estetyka słowa przechodzi przez językoznawczo-matematyczno-statystyczny filtr oprogramowania, tworząc na końcu nową estetykę – obrazu. Jeżeli jednak chodzi nam o badania naukowe, a nie o graficzne impresje, lepiej nie iść tą drogą, bo tu oczywiście kończą się wszelkie próby obiektywizmu

[36]T. Mizerkiewicz, prywatny e-mail z 13.07.2017.

badawczego, które legły kiedyś u podstaw badań ilościowych. Jeszcze w ubiegłym wieku pisał Edward Stachurski, że „zastosowanie metod statystycznych w językoznawczych i stylistycznych badaniach nad tekstem pozwala mieć pewność, że uzyskane wyniki opierają się na obiektywnych podstawach, niezależnych od subiektywnych osądów czytelnika”[37]. Wtóruje mu David Hoover: „Ilościowe metody badań literackich przedstawiają elementy i cechy tekstów literackich – w liczbach, stosując ścisłe i powszechnie stosowane wzory matematyczne, co umożliwia obiektywny pomiar, klasyfikację i analizę”[38]. Niejeden humanista cyfrowy przyznaje, że w świat komputerów pchnął go nihilizm poznawczy postmodernizmu, w którym jedynym prawdziwym stwierdzeniem ma być to, że jednej prawdy nie ma[39].

Nie przesadzajmy jednak z tym obiektywizmem – przed którym przestrzega też tekst Macieja Edera, towarzyszący „pierwszemu rzutowi oka na mapę literatury polskiej”[40]. To prawda, że pomiar i klasyfikacja dokonuje się w sposób, w którym subiektywne wybory badacza grają umiarkowaną rolę. Umiarkowaną, lecz wciąż widoczną, bo nawet najbardziej sumienny i bezstronny analityk musi podjąć kilka decyzji mocno obciążających jego sumienie. Jak duży korpus? Kiedy korpus można uznać za odpowiednio „reprezentacyjny”? Czy „reprezentacyjny” oznacza: uwzględniający różnice w liczbie dzieł różnych autorów – a więc dobrze, że Kraszewskiego jest w korpusie tak dużo, bo to nie jego wina, że Schulz zdążył napisać tak mało? A może właśnie należałoby zachować „sprawiedliwe”, bo egalitarystyczne proporcje? Z jednej strony niewspółmierność rozmiarów materiału jednego i drugiego zaburza językową równowagę tekstu – słowa częste z olbrzymiego dorobku Kraszewskiego (i z polskiego Szekspira) wywierają znacznie większy wpływ na wspólną dla wszystkich tekstów w korpusie listę niż schulzowskie perełki. Ale równocześnie taki właśnie jest pełny obraz literatury polskiej: Kraszewski, Jeż, Papi i Lem pisali dużo, inni znacznie mniej, i tego nie da się zmienić po śmierci pisarza czy pisarki – a często i za ich życia.

Drugim momentem nieobiektywnego wyboru jest etap wyboru parametrów analizy ilościowej. I znów: to prawda, że stylometria wciąż poszukuje tutaj konsensusu i wciąż wypracowuje metody, które mają ograniczyć wpływ takich czy innych ustawień programów analitycznych na otrzymane wyniki – i że jest to jeden z głównych, wspólnych celów tego środowiska naukowego[41]. Ale wątpliwości pozostają: czy na pewno uśrednianie wyników z wielu analiz jednostkowych prowadzi do uzyskania najbardziej stabilnych wyników? Czy raczej powinno się szukać tego jednego, jedynego a idealnego zestawu parametrów – najczęściej jest nim oczywiście liczba słów, których częstości się porównuje?

A potem następuje trzeci moment: ten, w którym wszystko zostało policzone, komputerowo przetworzone i wykreślone na płaszczyźnie we wszystkich barwach tęczy – i przychodzi humanista, i patrzy. Czy rzeczywiście widzi w tym gąszczu obiektywną prawdę, czy tylko dopasowuje własną wiedzę – i własną niewiedzę, bo przecież nie czytał tych, dajmy na to, dwóch i pół tysiąca tekstów – do kolorowych plam?

[37]E. Stachurski, *Słowa-klucze polskiej epiki romantycznej*, Kraków 1998, s. 11-12.

[38]D. Hoover, *Quantitative Analysis and Literary Studies*, [w:] *A Companion to Digital Literary Studies*, red. S. Schreibman, R. Siemens, Oxford 2007, s. 518.

[39]W. McCarty, *Getting There*..., s. 190.

[40]M. Eder, *Metody ścisłe w literaturoznawstwie*...

[41]Por. np. J. Rybicki, M. Eder, *Deeper Delta*...

Proponuję trochę mniej ambitny scenariusz: skoro nie wiemy tak naprawdę, jak językowe idiosynkratyzmy autorów przekładają się na statystykę słów, i to słów „synsemantycznych", i skoro nie mamy na razie jasnych teorii językoznawczych, które nam, humanistom-doświadczalnikom, służyłyby w taki sam sposób, w jaki fizyka teoretyczna wyznacza kierunki badaniom eksperymentalnym – korzystajmy z tego, co mamy. A mamy, po pierwsze, teksty metaliterackie i krytyczne; po drugie, rosnący materiał dowodowy, że analiza ilościowa często ukazuje relacje jak najbardziej zgodne z analizą jakościową. Skoro tak, to za każdym razem, gdy stylometria wskazuje na efekty niespodziewane, niezgodne z ową „jakościówką", może warto zastanowić się, czy nie jest to wskazówka do nowych odczytań. Najprostsze i najmniej kontrowersyjne zastosowanie stylometrii komputerowej – atrybucja autorska – zmienia przecież układ w polisystemie literackim za każdym razem, gdy wykrywa czy weryfikuje, kto naprawdę napisał konkretny tekst. Może warto rozciągnąć wiarę w analizę ilościową również na sytuacje, gdy analiza ilościowa przyciąga do siebie dwa teksty czy dwóch autorów, których podobieństwa nikt dotąd nie rozważał na podstawie lektury? W końcu żaden szanujący się lekarz nie postawi – często ratującej życie – diagnozy bez przejrzenia wyników analizy krwi i moczu. Stylometria oferuje badaczowi literatury takie właśnie badania laboratoryjne – może warto z nich czasem skorzystać?

SŁOWA KLUCZOWE:

MAKROANALIZA

Stylometria

CZYTANIE NA DYSTANS

Abstrakt:
W artykule przedstawiono wyniki analizy ilościowej najczęstszych słów korpusu ponad 2500 tekstów polskich: literatura polska od XIV do XXI wieku oraz polskie przekłady z angielskiego, francuskiego, rosyjskiego i (w mniejszym stopniu) innych języków. Wykazano istnienie w korpusie silnego sygnału rodzajowego i sygnału języka wyjściowego. Wyniki wskazują również na wyraźną odrębność stylometryczną języka polskich przekładów szekspirowskich i ich bardzo silne podobieństwo stylometryczne do polskiego dramatu romantycznego i neoromantycznego.

analiza wielowymiarowa

analiza sieciowa

literatura polska

przekłady polskie

Nota o Autorze:
Jan Rybicki (ur. 1963) – absolwent Instytutu Filologii Angielskiej UJ (1987). Wykładał na krakowskim Uniwersytecie Pedagogicznym (1991–2000, 2001–2011) oraz na Rice University w Houston (1996–1997, 2000–2001). Od roku 2011 jest adiunktem IFA UJ. Jego zainteresowania naukowe koncentrują się głównie na stylometrii komputerowej języka literackiego w oryginale i przekładzie. Oprócz artykułów naukowych udziela się jako tłumacz literatury anglojęzycznej – przetłumaczył już ok. 30 powieści takich pisarzy, jak Kingsley Amis, John le Carré, Douglas Coupland, William Golding, Nadine Gordimer, Francis Scott Fitzgerald, Kazuo Ishiguro, Kenzaburō Ōe czy Kurt Vonnegut. |