

Imiona i nazwisko

Maciej Jakub Eder

Wykształcenie, posiadane dyplomy

- | | |
|-----------|---|
| 1994–1999 | Studia magisterskie w Instytucie Filologii Polskiej Uniwersytetu Wrocławskiego; od roku 1995 w ramach indywidualnego toku pod kierunkiem prof. dr. Jerzego Woronczaka. |
| 1999 | Uzyskanie stopnia magistra filologii polskiej na podstawie pracy: <i>Andrzej Maksymilian Fredro, "Przysłowia mów potocznych" – projekt edycji krytycznej</i> (promotor: prof. dr Czesław Hernas). |
| 1995–1999 | Studia w Collegium Invisibile w zakresie folklorystyki (tutor: prof. dr Czesław Hernas). |
| 1999–2004 | Studia doktoranckie na Wydziale Filologicznym Uniwersytetu Wrocławskiego. |
| 2005 | Uzyskanie stopnia naukowego doktora nauk humanistycznych na podstawie pracy: <i>Moralistyka i polityka w "Przysłowiach mów potocznych" Andrzeja Maksymiliana Fredry. Studium o konwencjach gatunkowych i płaszczyznach lektury utworu</i> (promotor: dr hab. Marcin Cieński prof. UW; recenzenci: prof. dr Jacek Sokolski z UW, prof. dr Adam Karpiński z IBL PAN). |

Dotychczasowe zatrudnienie w jednostkach naukowych

- | | |
|-----------|--|
| 2004–2005 | Asystent w Instytucie Języka Polskiego PAN w Krakowie |
| 2006– | Adiunkt w Instytucie Języka Polskiego PAN w Krakowie |
| 2007– | Adiunkt w Instytucie Filologii Polskiej Uniwersytetu Pedagogicznego w Krakowie |

Wskazanie osiągnięcia wynikającego z art. 16 ust. 2 ustawy

Jednotematyczny cykl publikacji na temat stylometrii, tj. kwantytatywnej analizy różnicowania stylistycznego dzieł literackich. W skład cyklu wchodzi następujące publikacje (zostaną one omówione poniżej):

1. M. Eder, *Uwagi o rytmice pieśni „Jezusa Judasz sprzedał...” Władysława z Gielniowa – problem akcentu wyrazowego*, w: *Cantando cum citharista. W pięćsetlecie śmierci bł. Władysława z Gielniowa*, pod red. R. Mazurkiewicza, Warszawa 2006, s. 135–146.
2. A. Pawłowski, M. Krajewski, M. Eder, *Time series modelling in the analysis of Homeric verse*, „Eos” 97 (2010), s. 79–100.
3. M. Eder, *How rhythmical is hexameter: a statistical approach to Ancient epic poetry*, w: *Digital Humanities 2008: Book of Abstracts*, University of Oulu, Oulu 2008, s. 112–114.
4. J. Rybicki, M. Eder, *Deeper Delta across genres and languages: do we really need the most frequent words?*, „Literary and Linguistic Computing” 26 (2011), s. 315–321.

5. M. Eder, J. Rybicki, *Do birds of a feather really flock together, or how to choose test samples for authorship attribution*, „Literary and Linguistic Computing” 28 (2013), s. 229–236.
6. M. Eder, *Does size matter? Authorship attribution, short samples, big problem*, w: *Digital Humanities 2010: Conference Abstracts*, King's College London, London 2010, s. 132–135.
7. M. Eder, *Mind your corpus: systematic errors in authorship attribution*, „Literary and Linguistic Computing” 28 (2013), s. 603–614.
8. M. Eder, *Computational stylistics and Biblical translation: How reliable can a dendrogram be?*, w: *The Translator and the Computer*, red. T. Piotrowski, Ł. Grabowski, Wrocław 2013, s. 155–170.
9. M. Eder, J. Rybicki, *PCA, Delta, JGAAP and Polish poetry of the 16th and 17th centuries: who wrote the dirty stuff?*, w: *Digital Humanities 2009: Conference Abstracts*, University of Maryland, College Park (MD) 2009, s. 242–244.
10. M. Eder, J. Rybicki, *Stylometry with R*, w: *Digital Humanities 2011: Conference Abstracts*, Stanford University, Stanford (CA) 2011, s. 308–311.
11. M. Eder, *Bootstrapping Delta: a safety-net in open-set authorship attribution*, w: *Digital Humanities 2013: Conference Abstracts*, University of Nebraska-Lincoln, Lincoln (NE) 2013, s. 169–172.

Dorobek literaturoznawczy

Moje zainteresowania naukowe od wielu lat koncentrują się wokół zagadnień związanych z literaturą wieków XV–XVII, badanych zarówno w perspektywie literaturoznawczej, jak i językoznawczej, a także edytorskiej.

Za najważniejsze moje osiągnięcie literaturoznawcze uznałbym monografię zbioru aforystycznego *Przysłowia mów potocznych* autorstwa Andrzeja Maksymiliana Fredry (1620–1679), która jest rozszerzoną wersją mojej rozprawy doktorskiej¹. Książka omawia poczytny niegdyś, a dziś nieco zapomniany zbiorek celnych powiedzeń, który stanowił jeden z punktów węzłowych w kształtowaniu się polskiej aforystyki. Kolejne rozdziały książki zostały poświęcone kwestiom filologicznym, związkom Fredrowych „przysłów” – jak sam je nazywa – z przysłowiami tradycyjnymi, dalej kształtowaniu się nowego gatunku – aforyzmu, ostatnia zaś część rozprawy jest próbą lektury tekstu pod kątem poruszanej w nim problematyki.

Napisany wspólnie z dr Martą Wojtkowską-Maksymik monograficzny rozdział *Platońskie i neoplatońskie konteksty „humanitas” w literaturze polskiej XVI–XVIII wieku* stanowi próbę syntezy recepcji platonizmu w literaturze staropolskiej, ze szczególnym uwzględnieniem wątków antropologicznych². Rozdział z jednej strony podsumowuje dotychczasowe ustalenia kilku pokoleń badaczy, z drugiej zaś strony daje pogłębioną analizę kilku niedostrzeżonych kierunków recepcji (neo)platonizmu. Dr Wojtkowska-Maksymik jest autorką podrozdziałów *Godność i Miłość*, ja zaś zająłem się relacjami makrokosmosu (wszechświata) i mikrokosmosu (człowieka) w podrozdziale pt. *Harmonia mundi*.

¹ M. Eder, *U źródeł aforystyki polskiej. Studium o „Przysłowia...” Andrzeja Maksymiliana Fredry*, Wrocław 2008.

² M. Eder, M. Wojtkowska-Maksymik, *Platońskie i neoplatońskie konteksty humanitas w literaturze polskiej XVI–XVIII wieku*, w: *Humanitas. Projekty antropologii humanistycznej*, pod red. A. Nowickiej-Jeżowej, Warszawa 2010, t. 2, s. 93–159.

Szerzej rozumianym postawom polskiego humanizmu wobec literatury i kultury starożytnej grecko-rzymskiej poświęcone jest również studium *Obecność tekstów antycznych fundamentalnych dla humanizmu w kulturze staropolskiej: wspólnota czytelników*³, w którym starałem się zestawiać mało znane staropolskie przekłady dzieł Cyserona, Seneki, Waleriusza Maksymusa, Epikteta czy Plutarcha i pokazać ich funkcjonowanie jako zaplecza intelektualnego i kulturotwórczego w dawnej Polsce.

Funkcjonowaniem wątków platońskich i wspólnototwórczą rolą inspiracji antycznych zajmowałem się jako jeden z uczestników grantu MNiSW *Humanizm. Idee, nurty i paradygmaty humanistyczne w kulturze polskiej* (PBZ-MNiSW-03/II/2007), w częściach grantu kierowanych przez prof. dr Alinę Nowicką-Jeżową oraz dr. hab. Marcina Cieńskiego, prof. UWŹ).

Dorobek edytorski

Edytorstwo naukowe nie bez powodu nazywane bywa koroną filologii, korzysta bowiem zarówno z metod wypracowanych przez językoznawstwo, jak i z zaplecza literaturoznawczego, stając się niejako syntezą obu tych dyscyplin. Z tego też powodu za jeden z najważniejszych obszarów mojej działalności naukowej uważam edytorstwo tekstów dawnych.

Bez wątpienia najbardziej istotne jest tutaj opracowane wspólnie z prof. Romanem Mazurkiewiczem krytyczne wydanie traktatu Andrzeja Wolana *De libertate politica sive civili* (1572) wraz z polskim przekładem *O wolności Rzeczypospolitej albo ślacheckiej* (1606) pióra Stanisława Dubingowicza⁴. Andrzej Wolan był wybitnym teologiem, zwanym „papieżem kalwinistów litewskich”, a także pisarzem politycznym; jego rozprawa o wolności jest jednym z najważniejszych traktatów ustrojowych polsko-litewskiego renesansu. Przygotowane wydanie to w gruncie rzeczy dwie osobne edycje (oryginał i staropolski przekład): paralelne, lecz zaopatrzone w dwa niezależne aparaty krytyczne i z przeprowadzoną niezależnie krytyką tekstu. Prof. Mazurkiewicz wydał tekst polski, moim dziełem jest edycja łacińskiego oryginału. Rozbudowany komentarz towarzyszący edycji, odtwarzający kontekst kulturowy i literacki traktatu, jest dziełem wspólnym, przy czym prof. Mazurkiewicz zidentyfikował konteksty biblijne i nowożytne, ja zaś podjąłem się opracowania kontekstów starożytnych – greckich i łacińskich. Udział obu wydawców w przygotowaniu edycji był więc symetryczny, a ciężar prac rozłożony był równo po połowie. Edycja doczekała się recenzji i omówień krytycznych pióra W. Roczniaka, K. Daugirdasa, E. Patiejūniene i J. Niedźwiedzia⁵.

Znacznie mniejsze objętościowo jest wydanie łacińskich dokumentów dotyczących Stanisława Poklateckiego, pomniejszego pisarza jezuickiego z końca XVI wieku, zamieszczone jako kilkustronicowy aneks do wydania dwóch traktatów: *Snów dobrych obrona a szkodliwych przestroga* oraz *Pogrom czarnoksiężskie błędy, latawców zdrady i alchimiczne fałsze jako rozpłaska*⁶. Dokumenty dotyczące wstąpienia Poklateckiego do zakonu jezuitów, a następnie wystąpienia w okolicznościach uznanych za gorszące (nie znamy szczegółów) pochodzą z kilku rękopisów Archiwum Rzymskiego Towarzystwa Jezusowego oraz Biblioteki Uniwersyteckiej w Uppsali; oświetlają wiele dotychczas nieznanych szczegółów z życia pisarza.

Wśród przygotowanych przeze mnie wydań mieści się również garść tekstów zamieszczonych

³ M. Eder, *Obecność tekstów antycznych fundamentalnych dla humanizmu w kulturze staropolskiej: wspólnota czytelników*, w: *Humanizm polski i wspólnoty. Naród – społeczeństwo – państwo – Europa*, pod red. M. Cieńskiego, Warszawa 2011, s. 151–182.

⁴ A. Wolan, *De libertate politica sive civili / O wolności Rzeczypospolitej albo ślacheckiej*, wyd. M. Eder i R. Mazurkiewicz, red. W. Uruszczak, Warszawa 2010.

⁵ W. Roczniak, „Renaissance Quarterly” 65 (2012), s. 1172–1173; K. Daugirdas, „Archiv für Reformationsgeschichte. Literaturbericht” 40 (2011), s. 131; E. Patiejūniene, „Senoji Lietuvos Literatūra” 30 (2010), s. 329–344; J. Niedźwiedź, „Terminus” 24 (2011), s. 149–156.

⁶ S. Poklatecki, *O snach i czarach*, oprac. J. Krocak, E. Madeyska, aneksy oprac. M. Eder, Wrocław 2011, s. 183–188.

w witrynie Staropolska.pl – Jana Danieckiego *Przyjaciel szczery*, Justusa Lipsjusza *O stałości* w przekł. Janusza Piotrowicza (fragm.), Piotra Węzyka Widawskiego *O pchle, maluczkim ptaszku, elegija* (wyd. wspólnie z Radosławem Grześkowiakiem), a nade wszystko Jana Protasowicza *Konterfet człowieka starego*⁷. Przygotowane edycje, opatrzone komentarzem, są przymiarką do przygotowania większego zespołu tekstów staro- i średniopolskich w postaci korpusowej.

Na pograniczu zagadnień edytorskich i bibliologicznych znajdują się moje prace nad sporządzeniem katalogu polskich druków z XVI i XVII wieku znajdujących się w bibliotece zamku Skokloster, dokąd dotarły jako zdobycz wojenna z czasów „potopu” szwedzkiego, zrabowane przez generała Karla Gustawa Wrangla z którejś z wielkopolskich bibliotek. Katalog druków ze Skokloster, wydany w Sztokholmie i w wersji rozszerzonej w Warszawie⁸, notuje bez mała 300 pozycji, z czego stosunkowo duża część to druki unikatowe.

Dorobek językoznawczy

Leksykografia

Od roku 2004, jako pracownik Instytutu Języka Polskiego PAN, współuczestniczę w redagowaniu suplementu do wielotomowego *Słownika staropolskiego*. Mój udział trudno zmierzyć liczbą publikacji czy innymi wskaźnikami bibliometrycznymi, gdyż stanowi część (co prawda bardzo niewielką) rozpisanego na dziesięciolecie przedsięwzięcia leksykograficznego, które od samych swych początków zakładało wieloautorską i kolegialną pracę nad poszczególnymi hasłami. Pomijając zresztą sam efekt końcowy, czyli właściwe hasła słownikowe, praca w redakcji słownika historycznego obejmuje ekscerpcję źródeł, ustalanie datacji poszczególnych zabytków, odczytywanie kontekstów łacińskich z bardzo czasem uszkodzonych rękopisów *etc.* Odwołując się do konkretnego przykładu: w zbiorach Pracowni znajduje się spora liczba maszynopisów zawierających odczytania łacińskich rękopisów, sporządzone przez kilka pokoleń paleografów; do tego zbioru odczytów dołożyłem kilkadziesiąt dalszych stron – głównie z wczesnopiętnastowiecznych rękopisów Akt Grodzkich i Ziemskich przechowywanych w Archiwum Państwowym na Wawelu. W związku z tym liczba opracowanych przeze mnie artykułów hasłowych (ok. 200), które zostaną włączone do suplementu *Słownika staropolskiego*, nie oddaje w całości mojego rzeczywistego wkładu.

Bardziej wymiernym efektem moich prac przy *Słowniku* są opracowane wspólnie z prof. Wacławem Twardzikiem *Indeksy do Słownika staropolskiego*⁹, które doczekały się przychylnych recenzji¹⁰, a także dwa artykuły: jeden z nich omawia meandry XVII-wiecznej leksykografii polskiej¹¹, drugi zaś stara się zaproponować nowe odczytanie niejasnych fragmentów obsceniczej pieśni z 1416 roku, znanej jako *Cantilena inhonesta*¹² (współautorem tego tekstu jest ponownie prof. W. Twardzik).

⁷ Zob. <http://staropolska.pl>

⁸ M. Eder, *The „Polonica Collection” from Skokloster Castle*, Stockholm 2008 (Stockholm Slavic Papers, vol. 16); *Polonika ze zbiorów zamku Skokloster. Katalog*, oprac. M. Eder przy współud. E. Westin Berg, Warszawa 2008. Rec.: A. Znajomski, „Studia Źródłoznawcze” 48 (2011), w druku; E. Różycki, „Rocznik Biblioteki Narodowej” 42 (2011), s. 346–449; P. Buchwald-Pelcowa, „Rocznik Biblioteki Narodowej” 42 (2011), s. 350–354.

⁹ M. Eder, W. Twardzik, *Indeksy do Słownika staropolskiego: alfabetyczny, a tergo, verba absentia, verba expurgata*, Kraków 2007, ss. XVI+304.

¹⁰ T. Mika, „Język Polski” 88 (2008), s. 245–249; M.B. Majewska, „Poradnik Językowy” 5 (2008), s. 82–86; M. Kierkowicz, „Biuletyn Polskiego Towarzystwa Językoznawczego” 64 (2008), s. 277–283.

¹¹ M. Eder, *O słownikach osobliwych, czyli przyczynek do sarmackiej leksykografii*, w: *Amoenitates vel lepores philologiae*, pod red. R. Laskowskiego, R. Mazurkiewicza, Kraków 2007, s. 81–95.

¹² M. Eder, W. Twardzik, *Czy staropolska kicz/kić naprawdę była wyrazem nieprzyzwoitości? („Cantilena inhonesta” odczytana na nowo)*, w: *Od fonemu do tekstu. Prace dedykowane profesorowi Romanowi Laskowskiemu*, pod red. I. Bobrowskiego, K. Kowalik, Kraków 2006, s. 163–175.

Językoznawstwo kwantytatywne i stylometria

Przedstawiony poniżej pokrótce jednotematyczny cykl artykułów jest moim głównym osiągnięciem wskazanym jako podstawa habilitacji. Choć wszystkie omówione publikacje dotyczą analizy stylu wypowiedzi autorskiej z zastosowaniem metod statystycznych, dzielą się na dwa zasadnicze kręgi zainteresowań: (1) wersyfikacja (polska, grecka, łacińska), (2) atrybucja autorska.

Artykuł *Uwagi o rytmice pieśni „Jezusa Judasz sprzedał...” Władysława z Gielniowa*¹³ jest głosem w dyskusji o miejsce padania akcentu w XV-wiecznej polszczyźnie. Badacze są na ogół zgodni, że w wieku XIV w polszczyźnie utrzymywał się jeszcze akcent inicjalny, a w połowie XVI wieku już z pewnością paroksytoniczny; moment tej zmiany i jej dynamika pozostają niejasne. Podejście proponowane przeze mnie polega na porównaniu średniówek i klauzul pieśni Władysława z Gielniowa *Jezusa Judasz sprzedał* (1488) z jej czeskim pierwowzorem (akcent inicjalny) i z Psalmem 104 Jana Kochanowskiego (akcent paroksytoniczny). Przeprowadzona przeze mnie analiza pokazuje, że w wymowie Władysława z Gielniowa *nie istniał* wyraźny akcent inicjalny.

Opublikowany wspólnie z prof. Adamem Pawłowskim i dr. Markiem Krajewskim tekst *Time series modelling in the analysis of Homeric verse*¹⁴ jest próbą odpowiedzi na pytanie, czy w prozodii greckiego heksametru, opartej zasadniczo na iloczasiu, istniał również akcent dynamiczny, tzw. *ictus*. W artykule została postawiona hipoteza, że jeśli tekst grecki zakodowany jako sekwencja sylab akcentowanych i nieakcentowanych okaże się mocniej zrytmizowany niż ten sam tekst traktowany jako następstwo długich i krótkich sylab, będzie to przemawiało za rzeczywistym istnieniem iktu w epice greckiej. Analizie poddano kilkadziesiąt próbek z poematów Homerowych z zastosowaniem tzw. modelowania szeregów czasowych. Uzyskane wyniki wskazują na mocne zrytmizowanie obu cech prozodyjnych – zarówno iloczasu, jak i hipotetycznego iktu – co jest szczególnie ciekawe w zestawieniu z podobnym artykułem sprzed kilku lat, w którym została pokazana duża dysproporcja w zrytmizowaniu iloczynowym i przyciskowym heksametru łacińskiego¹⁵.

Niewielki artykuł *How rhythmical is hexameter?*¹⁶ jest rozwinięciem tez stawianych w poprzednich publikacjach; jego celem jest porównanie zrytmizowania łacińskich i greckich poematów epickich pochodzących z różnych epok i powstałych w odmienny sposób: w tradycji ustnej i w kulturze piśmiennej. Do badania ukrytych regularności w próbkach tekstowych wykorzystana została analiza Fourierska. Okazało się, po pierwsze, że istnieje fundamentalna różnica między rytmiką łacińską i grecką (heksametr łaciński nie wykazywał żadnych istotnych regularności, grecki dawał dość silny rytm), a po drugie, że w obrębie epiki greckiej dawała się zauważyć różnica między eposami archaicznymi i utworami powstałymi w późniejszych epokach.

Drugi krąg podejmowanych przeze mnie zagadnień dotyczy stylometrii, czyli statystycznej analizy stylu pisarskiego. Mniej więcej od czasów Augusta de Morgana (1806–1871) i Wincenego Lutosławskiego (1863–1954) stylometria stara się odpowiedzieć na pytanie, czy możliwa jest jednoznaczna identyfikacja autora dzieła literackiego (lub, w szerszym ujęciu, jakiegokolwiek tekstu pisanego) na podstawie analizy drobnych, dających się oszacować statystycznie różnic stylistycznych. Moje prace poświęcone są w szczególności problemom teoretycznym

¹³ M. Eder, *Uwagi o rytmice pieśni „Jezusa Judasz sprzedał...” Władysława z Gielniowa – problem akcentu wyrazowego*, w: *Cantando cum citharista. W pięćsetlecie śmierci bł. Władysława z Gielniowa*, pod red. R. Mazurkiewicza, Warszawa 2006, p. 135–146.

¹⁴ A. Pawłowski, M. Krajewski, M. Eder, *Time series modelling in the analysis of Homeric verse*, „Eos” 97 (2010), s. 79–100.

¹⁵ A. Pawłowski, M. Eder, *Quantity or stress? Sequential analysis of Latin prosody*, „Journal of Quantitative Linguistics” 8 (2001), s. 81–97.

¹⁶ M. Eder, *How rhythmical is hexameter: a statistical approach to Ancient epic poetry*, w: *Digital Humanities 2008: Book of Abstracts*, University of Oulu, Oulu 2008, s. 112–114.

i metodologicznym stylometrii, w mniejszym zaś stopniu skupiają się na rzeczywistej atrybucji poszczególnych dzieł literackich.

Od kilkudziesięciu mniej więcej lat daje się zauważyć intensywny rozwój metod stylometrycznych, przy czym uwaga badaczy skupia się na ogół na *efektywności* poszczególnych procedur statystycznych. Gdy miary bogactwa słownikowego okazały się nie dość dokładne i zanadto zależne od długości badanej próby, zwrócono się w stronę oszacowania średniej długości słów czy zdań i na tej podstawie wnioskowania o autorstwie; potem szukano innych miar różnicujących teksty. Wielkim krokiem naprzód było – i nadal jest – zastosowanie metod wielowymiarowych, czyli takich, które swe działanie opierają na jednoczesnym mierzeniu *wielu zmiennych*: na ogół są to frekwencje kilkudziesięciu lub kilkuset najczęstszych słów. Sukcesem okazały się analiza skupień, analiza głównych składowych, skalowanie wielowymiarowe, czyli techniki znane i od lat stosowane w naukach ścisłych i społecznych. Mimo że stylometria wypracowywała czasem swoje własne metody – przykładem Delta, Zeta oraz Iota opracowane przez Burrowsa¹⁷ albo zaproponowane przeze mnie dwie nieeuklidesowe miary odległości¹⁸ – prawie zawsze stosuje się techniki już istniejące, często operujące bardzo zaawansowanym aparatem matematycznym. Najbardziej spektakularne efekty daje użycie tzw. metod uczenia maszynowego, w tym najbliższych skurczonych centroid (nearest shrunken centroids)¹⁹ oraz maszyn wektorów nośnych (support vector machines)²⁰.

Ten kierunek rozwoju technik atrybucji autorskiej jest zupełnie naturalny. Zarazem jednak współczesna stylometria zdaje się nie doceniać faktu, że *wiarygodność* uzyskanych wyników jest równie ważna jak *efektywność* (precyzja) zapewniana przez zastosowanie najbardziej wyrafinowanych technik statystycznych. Chodzi mianowicie o to, że, przykładowo, sięgając po metodę wypracowaną dla filogenetyki molekularnej, językoznawca *ipso facto* zakłada (na ogół nieświadomie), że frekwencje słów w korpusie mają taką samą charakterystykę jak ekspresje genów w łańcuchu DNA, co niekoniecznie jest zgodne z prawdą. Tego typu zagadnień związanych z wiarygodnością eksperymentu stylometrycznego jest znacznie więcej; część z nich poruszam w kilku studiach napisanych samodzielnie lub z dr. Janem Rybickim, opublikowanych m.in. w czasopiśmie „Literary and Linguistic Computing” (czasopismo notowane jest na tzw. liście filadelfijskiej).

Artykuł *Deeper Delta across genres and languages*²¹ jest próbą odpowiedzi na pytanie, ile najczęstszych słów powinno być zmierzonych jednocześnie, by uzyskać optymalny wynik atrybucji autorskiej. W dotychczasowych badaniach jedni autorzy optowali za niewielką liczbą słów synsemantycznych, inni byli zdania, że tylko bardzo długie wektory zmiennych (tj. słów) dają wiarygodne wyniki: rozpiętość między poszczególnymi stanowiskami uczonych była ogromna. Współ z dr. Rybickim zaproponowaliśmy eksperyment złożony z tysięcy pojedynczych testów atrybucyjnych, testujących każdą możliwą kombinację częstych słów. Uzyskane wyniki sugerują jednoznacznie, że optymalne *a priori* warunki atrybucji nie istnieją i że występuje tu silna zależność od języka (w tym: stopnia fleksyjności), od gatunku i od innych czynników.

¹⁷ J. Burrows, ‘Delta’: a measure of stylistic difference and a guide to likely authorship, „Literary and Linguistic Computing” 17 (2002), s. 267–287; idem, All the way through: testing for authorship in different frequency strata, „Literary and Linguistic Computing” 22 (2007), s. 27–48.

¹⁸ „Eder’s Delta Distance” oraz „Eder’s Simple Distance” pokrótce opisane w: M. Eder, J. Rybicki, M. Kestemont, *Stylo: a package for stylometric analyses*, https://sites.google.com/site/computationalstylistics/stylo_howto.pdf

¹⁹ M. Jockers, D. Witten, C. Criddle, *Reassessing authorship of the ‘Book of Mormon’ using delta and nearest shrunken centroid classification*, „Literary and Linguistic Computing” 23 (2008), s. 465–491.

²⁰ M. Koppel, J. Schler, S. Argamon, *Computational methods in authorship attribution*, „Journal of the American Society for Information Science and Technology” 60 (2009), s. 9–26; E. Stamatatos, *A survey of modern authorship attribution methods*, „Journal of the American Society of Information Science and Technology” 60 (2009), s. 538–356.

²¹ J. Rybicki, M. Eder, *Deeper Delta across genres and languages: do we really need the most frequent words?*, „Literary and Linguistic Computing” 26 (2011), s. 315–321.

Studium *Do birds of a feather really flock together, or how to choose test samples for authorship attribution*²², choć dotyczy całkowicie innego aspektu eksperymentów atrybucyjnych, jest w pewnym sensie kontynuacją poprzedniego artykułu. W zaawansowanych metodach uczenia maszynowego stosuje się procedurę tzw. sprawdzianu krzyżowego (cross-validation). W największym skrócie: w kilku krokach losowo usuwa się pojedyncze teksty z korpusu i po chwili przywraca ponownie (chodzi o wprowadzenie czynnika losowego do eksperymentu), testując w ten sposób stabilność uzyskanych wyników. Standardowo wykonuje się 10 takich losowych zawirowań. Powyższą technikę językoznawcy przejęli *in extenso* z nauk ścisłych, nikt jednak nie badał jej rzeczywistej skuteczności na gruncie stylometrycznym. W podjętym z dr. Rybickim eksperymencie zastosowaliśmy 500 prób zamiast zwykłych 10, a w każdej próbie losowo dobieraliśmy *cały korpus* referencyjny zamiast żonglować pojedynczymi próbkami. Wyniki pokazały, że powszechnie stosowany 10-krotny sprawdzian krzyżowy jest rozwiązaniem dalece niedoskonałym: zmienność w obrębie korpusu tekstów okazała się znacznie większa niż uważano do tej pory.

W krótkim artykule *Does size matter? Authorship attribution, short samples, big problem*²³ opisuję eksperyment, który jest próbą rozwiązania jednego z najważniejszych problemów stylometrii: jaka jest minimalna długość tekstu, by atrybucja autorska była skuteczna i wiarygodna. Do problemu podszedłem w ten sposób, że z korpusu zawierającego stosunkowo długie teksty (m.in. powieści) wyłaniałem losowo, w bardzo dużej liczbie rund, coraz dłuższe „wirtualne” próbki tekstowe i te wyłonię losowe próbki poddawałem kontrolowanym testom atrybucyjnym. Zgodnie z przewidywaniami przy pewnej długości próbek następuje nasycenie statystyczne, tzn. efektywność atrybucji osiąga swe maksimum i pozostaje stabilna mimo dalszego zwiększania długości badanych tekstów. Zupełnie niespodziewany był natomiast fakt, że do osiągnięcia owego maksimum potrzeba aż 5000 słów dla większości korpusów (odpowiada to mniej więcej kilkunastu stronicom prozy) i tylko korpus łańciski dawał się skutecznie atrybuować przy próbkach długości ok. 2500 wyrazów.

Artykuł ten otrzymał nagrodę Paul Fortier Prize przyznaną przez Alliance of Digital Humanities Organizations (Londyn, czerwiec 2010) i niedługo ukaże się drukiem w pełnej wersji w jednym z renomowanych czasopism branżowych.

Zupełnie inny aspekt wiarygodności badań stylometrycznych podejmuję w tekście *Mind your corpus: systematic errors in authorship attribution*²⁴. Celem tego studium jest odpowiedź na pytanie, jak duży stopień nonszalancji w przygotowaniu korpusu jest akceptowalny, by test stylometryczny mógł być uznany za wiarygodny (np. jak duża liczba błędów literowych jest dopuszczalna, czy można pozostawić różne nieautorskie wstawki: przypisy, przedmowy, spisy treści itd.). Powyższe pytanie jest dość istotne w czasach łatwo pozyskiwalnych korpusów z różnych źródeł internetowych, gdy istnieje uzasadnione podejrzenie, że spora część tekstów dostępnych on-line jest niskiej lub bardzo niskiej jakości (nie mówiąc już o tym, że zaledwie drobny ich ułamek spełnia wymagania edycji krytycznej). Badania stylometryczne na ogół oznaczają przetwarzanie dużych zbiorów danych – czasem kilkudziesięciu czy kilkuset tekstów jednocześnie. Gdy czytanie każdego pliku tekstowego z osobna, porównywanie z wydaniem drukowanym i poprawianie błędów nie wchodzi w rachubę, pojawia się pilna potrzeba automatycznego oszacowania ich wartości. W omawianym artykule przeprowadzam symulację „brudnego” korpusu (czyli zawierającego liczne błędy literowe), ale także omawiam inne rodzaje niechcianego szumu pozaautorskiego, w tym zmiany wprowadzone przez kopistów, interpolacje, warianty ortograficzne, a także poprawki pochodzące od współczesnych edytorów. Jeden z eks-

²² M. Eder, J. Rybicki, *Do birds of a feather really flock together, or how to choose test samples for authorship attribution*, „Literary and Linguistic Computing” 28 (2013), s. 229–236.

²³ M. Eder, *Does size matter? Authorship attribution, short samples, big problem*, w: *Digital Humanities 2010: Conference Abstracts*, King's College London, London 2010, s. 132–135.

²⁴ M. Eder, *Mind your corpus: systematic errors in authorship attribution*, „Literary and Linguistic Computing” 28 (2013), s. 603–614.

perymentów został poświęcony weryfikacji założenia leżącego u podstaw wszystkich bez mała studiów atrybucyjnych: chodzi mianowicie o milczące (błędne) założenie, że tekst literacki pisany jest ręką jednego autora i nie ma w nim śladów innych rąk autorskich, gdy tymczasem każdy autor wyrasta z pewnej tradycji literackiej, czerpie – pośrednio lub bezpośrednio, świadomie lub nieświadomie – ze swoich poprzedników i z całego dorobku literackiego (oraz językowego!) poprzednich pokoleń. Eksperyment, o którym mowa, polegał na symulacji intertekstualnych wpływów pomiędzy badanymi tekstami; celem było empiryczne sprawdzenie, jak bardzo można rozwodnić tekst wpływami obcymi, by stylometryczny profil autorski wciąż był w nim widoczny. Uzyskane wyniki były nader ciekawe, szczególnie dla korpusu łacińskiego: okazało się, że usuwając nawet 40% losowo wybranych słów z tekstu łacińskiego i zastępując je taką samą liczbą dowolnych innych słów łacińskich, otrzymujemy nadal bardzo stabilne wyniki i bardzo wyrazisty sygnał autorski (!).

Kolejny artykuł stara się odpowiedzieć na pytanie, które z tzw. znaczników stylu (style-markers) zawierają najsilniejszy odcisk palca autorskiego²⁵. Tekst ten jest głosem w długiej debacie na ten temat. Od lat mniej więcej 60. XX wieku dość powszechnie uważa się, że frekwencje najczęstszych słów mają największą siłę dyskryminacyjną, z czasem jednak proponowano inne znaczniki: dłuższe lub krótsze sekwencje liter, najczęstsze zbitki słów (np. szeregi dwóch sąsiadujących słów), a także sekwencje kodów gramatycznych (np. części mowy). Mój głos w dyskusji, oparty na szeregu kontrolowanych eksperymentów – czyli takich, w których badacz zna autorstwo badanych tekstów literackich i sprawdza, jak dużą część z nich komputer przypisał właściwemu autorowi – miał na celu porównanie dokładności poszczególnych znaczników stylu, ale – niejako przy okazji – pokazał fundamentalną różnicę między analitycznym językiem angielskim i silnie fleksyjnym łacińskim w działaniu par wyrazowych (tzw. bi-gramów słownych) jako znacznika stylu.

W następnym studium²⁶ przeprowadzone zostało porównanie dwóch wersji Nowego Testamentu – greckiego oryginału i jego łacińskiego przekładu znanego jako Wulgata – za pomocą metod stylometrycznych. Mimo że artykuł omawia kwestie związane ze zróżnicowaniem stylometrycznym poszczególnych ksiąg biblijnych, jego głównym celem jest zmierzenie się z problemem wiarygodności w stylometrii, szczególnie gdy używane są tzw. metody nienadzorowane. Zaproponowany został stosunkowo prosty sposób weryfikacji wiarygodności wykresów analizy skupień przez wielokrotne powtórzenie wyjściowego testu, lecz z każdorazową permutacją danych wejściowych.

Cykl publikacji stylometrycznych zamykają trzy krótkie artykuły: jeden to studium przypadku, w którym kilka popularnych metod analizy wielowymiarowej zostało zastosowanych do ustalenia autorstwa cyklu erotyków przypisywanych Mikołajowi Sępowi Szarzyńskiemu²⁷, drugi jest prezentacją programu komputerowego przeznaczonego do analiz stylometrycznych²⁸ (por. więcej na ten temat poniżej), trzeci wreszcie stanowi próbę rozwiązania najtrudniejszego problemu metodologicznego atrybucji autorskiej, czyli tzw. problemu zbioru otwartego (open-

²⁵ M. Eder, *Style-markers in authorship attribution: a cross-language study of the authorial fingerprint*, „Studies in Polish Linguistics” 6 (2011), s. 99–114.

²⁶ M. Eder, *Computational stylistics and Biblical translation: How reliable can a dendrogram be?*, w: *The Translator and the Computer*, red. T. Piotrowski, Ł. Grabowski, Wrocław 2013, s. 155–170.

²⁷ M. Eder, J. Rybicki, *PCA, Delta, JGAAP and Polish poetry of the 16th and 17th centuries: who wrote the dirty stuff?*, w: *Digital Humanities 2009: Conference Abstracts*, University of Maryland, College Park (MD) 2009, s. 242–244.

²⁸ M. Eder, J. Rybicki, *Stylometry with R*, w: *Digital Humanities 2011: Conference Abstracts*, Stanford University, Stanford (CA) 2011, s. 308–311. Por. też: M. Eder, M. Kestemont, J. Rybicki, *Stylometry with R: a suite of tools*, w: *Digital Humanities 2013: Conference Abstracts*, University of Nebraska-Lincoln, Lincoln (NE) 2013, s. 487–489.

set problem)²⁹. Tekst ten, ze względu na ważność stawianych w nim tez, jest jedynie zarysem pogłębionego studium, nad którym obecnie pracuję.

Chcąc ustalić autorstwo anonimowego tekstu literackiego (czy jakiegokolwiek innego), stylometra stoi zasadniczo przed zadaniem zebrania jak największej liczby tekstów napisanych przez możliwych „kandydatów”, czyli pisarzy, którzy mogli byli napisać ów anonimowy tekst. (Na ogół dodaje się do korpusu również kilka innych tekstów z epoki, które służą jako grupa kontrolna). Bez względu na to, jak bardzo wysublimowana metoda zostanie użyta, test stylometryczny zawsze polega na znalezieniu „najbliższego sąsiada” (nearest neighbor), a zatem najbliższego stylometrycznie tekstu spośród wszystkich zebranych w korpusie. Problem zbioru otwartego polega na tym, że badacz nigdy nie może mieć całkowitej pewności, że korpus referencyjny zawiera próbki wszystkich kandydatów do autorstwa. Czasem zaś mamy do czynienia ze zbiorem otwartym *ex definitione*: na przykład jeśli przeprowadzimy analizę porównawczą anonimowej *Batrachomyomachii* z innymi zachowanymi eposami greckimi: Homera, Hezjoda, Apolloniosa, Aratosa i Nonnosa, to któryś z tych poetów zostanie wskazany jako najbardziej prawdopodobny autor *Batrachomyomachii*, ale takie wskazanie będzie z gruntu błędne (ewentualnego autorstwa Pigresa z Helikarnasu sprawdzić nie sposób, bo po prostu brak materiału porównawczego).

Do tej pory nie udało się znaleźć satysfakcjonującego rozwikłania tego problemu. Próby rozwiązania idą w kilku kierunkach³⁰; moja propozycja zakłada wprowadzenie do korpusu kontrolowanego szumu i testowaniu stabilności uzyskanych wyników. W dużej liczbie iteracji warunki eksperymentu zostają losowo zmienione, a uzyskane wyniki porównane między sobą. Oczywiście, wprowadzenie elementu losowego sprawia, że wyniki są zawsze rozbieżne, ale stopień tej rozbieżności pozwala wiarygodnie ocenić podobieństwo badanych tekstów. Jeśli w korpusie znajdują się teksty rzeczywiście podobne do siebie (np. *Pride and Prejudice* Jane Austen oraz *Emma* tej samej autorki), to nawet symulowane trzęsienie ziemi nie zburzy tego podobieństwa, jeśli natomiast mamy do czynienia ze „zbiorem otwartym”, to dodany szum natychmiast sprawia, że wyniki stają się chaotyczne i wysoce niejednoznaczne. Ta właśnie *niejednoznaczność* jest dla badacza sygnałem ostrzegawczym, że korpus referencyjny jest niekompletny.

Moja metoda, mimo że wciąż jest na etapie testów i dopracowywania szczegółów matematycznych, zapewnia skuteczność porównywalną z nowoczesnymi technikami atrybucyjnymi, ale zarazem jest bez porównania skuteczniejsza w ostrzeganiu przed błędem „zbioru otwartego”.

Działalność na rzecz społeczności akademickiej

Od kilku lat jestem dość często proszony o recenzowanie artykułów nadsyłanych do czasopisma „Literary and Linguistic Computing” (lista filadelfijska, impact factor 0,717), a także „International Journal of Corpus Linguistics” (lista filadelfijska, impact factor 0,592); oprócz tego regularnie recenzuję tzw. rozszerzone streszczenia nadsyłane na doroczną konferencję „Digital Humanities” (konferencja ta ma bardzo restrykcyjny system recenzji, sprawiający, że zaledwie 30–40% zgłoszeń zostaje zaakceptowanych). W zeszłym roku otrzymałem z Katolickiego Uniwersytetu w Lowanium prośbę o ocenę dorobku jednego z kandydatów biorących udział w konkursie na stanowisko profesora na tamtejszym Wydziale Filologicznym.

²⁹ M. Eder, *Bootstrapping Delta: a safety-net in open-set authorship attribution*, w: *Digital Humanities 2013: Conference Abstracts*, University of Nebraska-Lincoln, Lincoln (NE) 2013, s. 169–172.

³⁰ Koppel, Schler, Argamon, *op. cit.*; B. Schaalje, P. Fields, M. Roper, G. Snow, *Extended nearest shrunken centroid classification: A new method for open-set authorship attribution of texts of varying sizes*, „Literary and Linguistic Computing” 26 (2011), s. 71–88; B. Schaalje, N.J. Blades, T. Funai, *An open-set size-adjusted Bayesian classifier for authorship attribution*, „Journal of the American Society for Information Science and Technology” 2013, doi: 10.1002/asi; W. Scheirer, A. Rocha, A. Sapkota, T.E. Boulton, *Towards open set recognition*, „IEEE Transactions on Pattern Analysis and Machine Intelligence” 10 (2012), s. 1–17.

Innego rodzaju wkładem intelektualnym jest rozwijane przeze mnie oprogramowanie, przekazywane społeczności akademickiej na licencji wolnego kodu źródłowego (tzw. open source). Nieuchwytnie dla indeksów cytowań i innych wskaźników bibliometrycznych, owe programy są instalowane i używane z pewnością znacznie częściej, niż czytane są moje rozprawy naukowe.

Jedną z owych kontrybucji jest rozszerzenie języka składu komputerowego $\text{\LaTeX}$ – w pakiecie “microtype” dopisałem kod odpowiedzialny za obsługę alfabetów środkowoeuropejskich. Pakiet “microtype” służy do tworzenia wysokiej klasy składu komputerowego, w którym stosuje się wyrównanie optyczne marginesów oraz automatyczne skalowanie glifów (nie wchodząc w szczególności: celem jest zharmonizowanie kolumny tekstu przez zastosowanie kilku niezauważalnych gołym okiem zabiegów typograficznych). Pakiet ten, wraz moimi dodatkami, dostępny jest domyślnie we wszystkich wersjach systemu $\text{\LaTeX}$ zainstalowanych na świecie.

Przedsięwzięciem znacznie ambitniejszym programistycznie jest stworzone przeze mnie i ciągle rozwijane, przy współudziale dr. Jana Rybickiego i dr. Mike’a Kestemonta, rozszerzenie języka R, mianowicie pakiet “stylo”. Jest to kompletna biblioteka programistyczna, oferująca zestaw niskopoziomowych funkcji-komend do różnych operacji związanych z przetwarzaniem tekstu, będąca zarazem gotowym, względnie łatwym w użyciu programem (zestawem funkcji wysokopoziomowych) do różnego rodzaju analiz stylometrycznych. Pakiet ten jest od pewnego czasu dostępny w repozytorium CRAN (<http://cran.r-project.org>), a zatem wchodzi w skład oficjalnej dystrybucji środowiska programistycznego R. W zeszłym roku pakiet ten został nominowany do nagrody “Digital Humanities Awards” (<http://dhawards.org/dhawards2012/nominations>).

Przy okazji pracy nad pakietem “stylo” powołałem, wspólnie z dr. Janem Rybickim, nieformalny zespół badawczy Computational Stylistics Group (<https://sites.google.com/site/computationalstylistics/>), którego jednym z celów jest dalszy rozwój metod i narzędzi stylometrycznych.

Działalność międzynarodowa

Po uzyskaniu stopnia doktora uczestniczyłem w kilku konferencjach międzynarodowych: w Waszyngtonie (Uniwersytet Stanu Maryland), Palo Alto (Uniwersytet Stanforda), w Nowym Jorku (Uniwersytet Nowojorski), w Lincoln (Uniwersytet Stanu Nebraska), Londynie (King’s College London), Hamburgu, Oulu, Belgradzie, Würzburgu i Nijmegen.

Na zaproszenie różnych instytucji zagranicznych prowadziłem wykłady w Sztokholmie, w Palo Alto (Uniwersytet Stanforda), w Hadze (Instytut Huygensa Królewskiej Niderlandzkiej Akademii Nauk), w Amsterdamie (Uniwersytet Amsterdamski oraz Instytut Meertensa Królewskiej Niderlandzkiej Akademii Nauk), w Groningen i w Padwie.

Samodzielnie lub wspólnie z dr. Janem Rybickim prowadziłem warsztaty poświęcone metodom kwantytatywnym w badaniach języka i literatury: w Lipsku (trzykrotnie), w Hamburgu, w Getyndze i w Hadze (w sumie ponad 100 godzin dydaktycznych).

Do innego rodzaju wyjazdów i staży zagranicznych zaliczyć można miesięczną kwerendę w bibliotekach londyńskich (stypendium Fundacji Lanckorońskich z Brzezia), trzytygodniową kwerendę w bibliotece Zamku Skokloster, bibliotece katedry w Strangrass oraz w sztokholmskim Riksarkivet (stypendium Fundacji Ingvara Anderssona), dwutygodniowy staż naukowy na Uniwersytecie w Groningen (w Center for Language and Cognition), współpracę przy projekcie „The Riddle of Literary Quality” (dwutygodniowy pobyt w Hadze, w Instytucie Huygensa Królewskiej Niderlandzkiej Akademii Nauk), wreszcie trzymiesięczny pobyt w Amsterdamie jako stypendysta eHumanities Center przy Królewskiej Niderlandzkiej Akademii Nauk (tzw. visiting fellowship).

