ADRIAN TRZOSS

ORCID: 0000-0002-6287-418X

Komputerowo wspomagane metody badania tekstów w polskiej perspektywie

Computer aided methods of text research in the Polish perspective

Streszczenie: W niniejszym artykule autor przedstawia stosowanie metod analizy przetwarzania języka naturalnego (NLP) w obszarze polskich badań. W analizie uwzględniono trzy pola badawcze: socjologiczne, politologiczne oraz literaturoznawcze. Omówione zostały prace takich badaczy, jak Marek Troszyński, Paweł Matuszewski oraz Maciej Eder. Efektem przeprowadzonej analizy było nakreślenie najważniejszych aspektów metodologicznych związanych z używaniem metody NLP: kontekstów, możliwości oraz zagrożeń. Finalnie wskazano dalsze perspektywy badawcze, w których stosowanie omawianych metod może przynieść potencjalnie pozytywne rezultaty.

Słowa kluczowe: NLP, stylometria, język, analiza sieciowa

Summary: In the following paper author discuss the natural language processing method (NLP) usage in polish academic literature. In the analysis three fields were pointed out: sociology, political science and literature science. Three groups of texts were presented from Marek Troszyński, Paweł Matuszewski and Maciej Eder. As the result of the conducted analysis author emphasized the most important methodological aspects of NLP usage: contexts, opportunities and risks. Finally, author indicated areas for the further research where NLP would be beneficial method.

Keywords: NLP, stylometry, language, social network analysis

ADRIAN TRZOSS

Analiza języka naturalnego w perspektywie polskiej

Niniejszy artykuł jest pokłosiem analizy funkcjonowania metod badania tekstów w polskiej myśli naukowej. Nie jest niczym odkrywczym stwierdzenie, iż komputeryzacja i digitalizacja pozwoliły akademikom na spojrzenie z nowej perspektywy na stare problemy, czy też rozwiązania nowych. Kolejnym atutem zdaje się być przyspieszenie prac nad materiałami źródłowymi, nie tylko dzięki upowszechnieniu cyfrowych wersji tekstów, ale także dzięki zautomatyzowaniu wcześniej manualnych czynności (jak na przykład obliczenia). Implementacja metod znanych z nauk ścisłych nie odbywa się jednak bezrefleksyjnie, ani też bez stosownych modyfikacji do specyfiki nauk społecznych oraz humanistyki. W niniejszej pracy zamierzam przedstawić trzy polskie kręgi badawcze, w których autorzy stosowali komputerowo wspomagane metody w pracach nad tekstami źródłowymi. Efektem przeprowadzonej analizy stosowanej przez nich metodologii będzie wskazanie zalet i wad implementowanych podejść, a także próba oceny ich przydatności w innych obszarach badawczych. Nie jest natomiast założeniem walidacja wyników, czy interpretacji badań wymienionych niżej autorów, ani tym bardziej krytyka ich podejścia metodologicznego.

Wybór polskich obszarów badawczych wiąże się z poczuciem niedostatku literatury polskojęzycznej kierowanej do krajowego odbiorcy[136]. Nie sposób jednoznacznie wskazać na przyczynę takiego stanu rzeczy, niemniej jednak pewnych poszlak dostarcza tekst Macieja Edera, w którym przestrzega on przed pułapkami mitycznego paradygmatu obiektywizmu związanego z implementacją metod nauk ścisłych do badań literaturoznawczych[137]. Humanistyka mocno osadzona w badaniach jakościowych i dyskursach nad interpretacją wyników opiera się w mniejszym stopniu na „twardych" danych otrzymanych drogą, możliwych do walidacji i powtórzenia, analiz ilościowych. Niemniej, jak wykazane zostanie dalej, sięgnięcie po takie metody nie musi oznaczać konkurencyjnego czy „lepszego" podejścia do tradycyjnych problemów badawczych humanistyki.

Postawiony cel badawczy wiąże się z ewaluacją funkcjonowania komputerowych metod analizy tekstów źródłowych. Trzy obszary, choć interdyscyplinarne, można określić jako badania socjologiczne, politologiczne oraz literaturoznawcze (a nawet idąc o krok dalej – językoznawcze). Wszystkie trzy, choć operujące innymi tradycjami i problemami badawczymi, czerpią ze wspólnego rdzenia metod, jakim jest w wersji bazowej przetwarzanie języka naturalnego (NLP, *natural language processing*)[138]. Jest to zbiór metod wpisujący się w językoznawstwo i lingwistykę komputerową stosowany do tekstów zdigitalizowanych, a także powstałych pierwotnie w wersji cyfrowej. Najczęściej tekst zamienia się w dogodną do obliczeń

[136] W. Babik, *Język naturalny w wyszukiwaniu informacji i problemy jego przetwarzania*, „Zagadnienia Informacji Naukowej" 2013, nr 1, s. 37–47; P. Malak, *Rozwój badań nad przetwarzaniem języka naturalnego*, „Zagadnienia Informacji Naukowej" 2010, nr 2, s. 21–30; M. Eder, *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii*, „Teksty Drugie" 2014, nr 2, s. 90–105.

[137] M. Eder, *Metody ścisłe*…

[138] P. Malak, *Rozwój badań*...

reprezentację liczbową, na przykład dokonując wektoryzacji słów, tj. zamieniając je w wektory w przestrzeni wielowymiarowej grupując względem zbieżności semantycznej (przykładowo słowa „król" i „królowa" będą w przestrzeni wielowymiarowej wektorami w bliskiej odległości, zaś wyraz „kot" w dalszej od nich)[139]. Jest to oczywiście jedna z wielu możliwości, o czym dalej.

To, co przesądziło o wyborze poniższych tekstów, to teoretyczna uniwersalność zastosowanych w nich metod. W różnych obszarach badawczych, pracujących z tekstem, można z sukcesami stosować algorytmy do analiz statystycznych. W części socjologicznej omówione zostaną badania Marka Troszyńskiego, który analizował narracje w polskojęzycznym *Webie*, dotyczące mniejszości ukraińskiej. Jest to interesujący przykład współpracy między tradycyjnymi badaniami „niekomputerowymi" nad sentymentem (negatywnym, a także mową nienawiści, *hate speech*) a algorytmami uczenia maszynowego, tj. łącząc pracę manualną badaczy z analogicznymi zabiegami programu. W drugim przypadku przedstawione zostały prace Pawła Matuszewskiego, który zbadał funkcjonowanie mediów i partii politycznych w mediach społecznościowych. Dzięki ewaluacji stosowanych algorytmów był w stanie w dużej mierze dokonać walidacji funkcjonującej hipotezy o tzw. komorach pogłosowych (*echo chamber*) oraz powiązanych z nimi zjawiskach polaryzacji społecznej. Badania te zostały uzupełnione o pochodne od warstwy tekstowej metadane im towarzyszące (polubienia, statystyki komentarzy itd.). W ostatniej części zaprezentowane będą badania Macieja Edera, który wskazuje nie tylko na możliwości jakie dają metody komputerowe, ale także zwraca uwagę na istotny problem pozornego obiektywizmu i ostrożności w ich stosowaniu. Posiłkując się analizami NLP, Eder dokonuje potwierdzenia hipotezy profesora Tomasza Jasińskiego o powiązaniu między „Mnichem z Lido" a „Gallem Anonimem". Interesującym jest tu fakt, iż stosowanie algorytmów pozwoliło na znaczny postęp w badaniach tradycyjnych nad dyskutowanym już od dekad problemem ustalenia autorstwa tekstów średniowiecznych. Wartością dodatnią w rozważaniach Edera jest też szczegółowe omówienie kontekstu stosowanych metod oraz zagrożeń i potencjalnych pułapek z nimi związanych. W ostatniej części zostanie wskazany możliwy dalszy kierunek badań z zastosowaniem omawianych metod, zarówno wobec źródeł „tradycyjnych" (tj. nie-cyfrowych), jak i tych współcześnie powstających (*born-digital*, m.in. materiałów z mediów społecznościowych).

Ujęcie socjologiczne. *Natural language processing* w badaniach nad sentymentem a pół-automatyzacja badań

Marek Troszyński jest autorem i współautorem prac poświęconych wykorzystaniu automatycznego kodowania i klasyfikowania materiałów w badaniach so-

[139] R. Bartusiak, Ł. Augustyniak, T. Kajdanowicz, P. Kazienko, M. Piasecki, *WordNet2Vec: Corpora agnostic word vectorization metod*, „Neurocomputing" 2019, nr 326–327, s. 141–150.

cjologicznych[140]. W badaniach autor stosuje metody z początku mieszane – manualne, półautomatyczne i automatyczne, prezentując otrzymane wyniki. W artykule *Analiza treści witryn internetowych z wykorzystaniem automatycznego kodowania*[141] wskazuje na istotność poprawnego przygotowania arkusza kodowego, jako podstawy do dalszej pracy socjologa. Za jego pomocą, do konkretnych partii tekstu zostanie przypisany odpowiedni kod, np. kategoria tematyczna, jak sport czy polityka (w badaniu zaprojektowano 11 kategorii). Koncepcja półautomatycznego kodowania polegała na przygotowaniu algorytmu uczenia maszynowego (*machine learning*), który w oparciu o przygotowany korpus treningowy „nauczy się" odpowiednio kodować losowe teksty. Korpus treningowy, czyli zbiór tekstów z przypisanymi kodami, służy programowi do rozpoznawania, jakie elementy tekstowe są charakterystyczne w danej kategorii (np. słowo „poseł" w kategorii polityka; „piłkarz" w kategorii sport itd.). Z całej bazy źródłowej (3700 adresów URL) wyselekcjonowano 1600 fragmentów, które następnie zostały manualnie zakodowane przez zespół koderów (6 osób), a stanowiły korpus treningowy. Następnie program, na podstawie wniosków wyciągniętych z analizy korpusu treningowego, dokonał automatycznego kodowania. Uśrednione wyniki kompletności i dokładności zastosowanego algorytmu, dla 10 prób, okazały się odpowiednio między 73 a 100 oraz 24 a 95 punktów. Jak sam autor badania zaznacza: „Te dane nie pozwalają nam traktować wyników kodowania automatycznego jako miarodajnych dla analizowanych tekstów. Ten typ danych stanowi pierwsze oszacowanie, przybliżenie, które ze względu na bardzo dużą liczbę tekstów, nie było możliwe do uzyskania tradycyjnymi metodami"[142].

Dalsze próby półautomatycznego kodowania w badaniach socjologicznych podjął w dwuczęściowym raporcie[143]. Tematyka raportu dotyczyła stosunku Polaków wobec Ukraińców w perspektywie stosowanego sentymentu w treściach internetowych (media społecznościowe). W oparciu o zewnętrzną firmę Sentione, pozyskano 1,2 mln tekstów zawierających słowa kluczowe: „Ukraina, Ukrainiec, Ukrainka, Ukraińcy, Ukrainki". Zebrane dane pochodziły z okresu od października 2016 do grudnia 2017 roku, gdzie, jak sam zaznacza, ze względu na testowanie narzędzia monitoringu treści internetowych, realny zakres czasowy źródeł to grudzień 2016 i listopad 2017 roku. Mimo oparcia się w swoim badaniu na teoriach poświęconych mechanice mediów społecznościowych oraz analizie ję-

[140] M. Troszyński, *Ukraina i Ukraińcy w polskim dyskursie internetowym. Analiza jakościowo-ilościowa tekstów zamieszczanych w mediach społecznościowych* [w:] *Raport. Mniejszość ukraińska i migranci z Ukrainy w Polsce. Analiza dyskursu*, red. P. Tyma, Warszawa 2018, s. 103–185; M. Troszczyński, *Ukraina i Ukraińcy w polskich mediach społecznościowych* [w:] *Raport 2. Mniejszość ukraińska i migranci z Ukrainy w Polsce. Analiza dyskursu*, red. P. Tyma, Warszawa 2019, s. 49–76; M. Troszyński, *Analiza treści witryn internetowych z wykorzystaniem automatycznego kodowania* [w:] *Metody badań online*, red. P. Siuda, Gdańsk 2016, s. 83–103.

[141] Tenże, *Analiza treści witryn…*, s. 92–95.

[142] Tamże.

[143] M. Troszyński, *Ukraina i Ukraińcy w polskim dyskursie internetowym…*; tenże, *Ukraina i Ukraińcy w polskich mediach społecznościowych…*

zyka naturalnego zdecydowano się na dosyć zawężony i dyskusyjny dobór słów kluczy. W badaniu nie zaznaczono problemów, z którymi boryka się nauka podczas stosowania analiz NLP tj. kontekstowego rozumienia treści, ironii, skrótów myślowych czy np. tekstów odnoszących się do badanej społeczności, które nie zawierają określeń wskazanych w słowach kluczach (np. wyrażeń potocznych czy dysfemizmów, które mogą być uważane za obraźliwe m.in. „banderowcy" albo „Ukry"). Nie wskazano także w omawianym tekście na zawiłości metodologiczne, takie jak problem zakresu badanego materiału. Mimo, iż w drugiej części raportu[144] zaznacza, iż do korpusu tekstowego nie włączono tych, które traktowały o Ukraińcach jako poszczególnych osobach (obszarem zainteresowania był stosunek do narodu ukraińskiego), to jednak nie otrzymujemy informacji na jakich zasadach to zostało dokonane. Problematyczna jest także kwestia postrzegania narodu poprzez jego jednostki tj. czy recepcja osoby nie prowadzi do uogólnienia na cały naród i *vice versa*. Problem kontekstu wyszukiwanych materiałów badawczych i filtracja względem interesujących tematów/haseł/słów kluczy zostały również zaobserwowane przez autora niniejszego tekstu. W trakcie badań nad ostatnimi tygodniami kampanii przedreferendalnej „Brexit" poddano analizie wpisy na portalu Facebook Davida Camerona (ówczesny premier) oraz Nigela Farage'a (czołowy zwolennik Brexitu). W przytoczonym artykule wskazano na problem rozróżnienia, które wpisy dotyczyły ogólnie polityki wewnętrznej Wielkiej Brytanii, a które stricte odnosiły się do trwającej kampanii[145].

Kolejną kwestią metodologiczną do przemyślenia jest porównanie skuteczności klasyfikowania wpisów przez następujące podmioty: wolontariuszy, zespół badawczy oraz algorytm (wybrano metodę fastText[146]). W zakresie poprawnego przypisania wolontariusze uzyskali 54% i 36% skuteczności (odpowiednio negatywność i przypisanie do obszaru tematycznego), a zespół 63% i 40%. Na podstawie zakodowanego przez nich korpusu algorytm został „wyuczony" z precyzją 68% dla negatywności i 44% dla rozpoznawania obszarów tematycznych. Jak zauważono wcześniej, w metodzie kodowania półautomatycznej bardzo istotnym jest poprawne przygotowanie korpusu treningowego. Jednak to, co może budzić pewne obawy, to wyciąganie ogólnych wniosków dla dużego zbioru materiałów, którego skuteczność jest przedstawiona powyżej. Innymi słowy, jeśli cztery na dziesięć tekstów zostały odpowiednio oznaczone pod kątem negatywności treści wypowiedzi, to wyciąganie na tej podstawie kolejnych wniosków może nie mieć pokrycia w rzeczywistości. Co więcej, w drugiej części raportu[147] zaznaczono, iż algorytm korzystał z innego korpusu treningowego, jakim były opinie pozyskane z forów takich jak TripAdvisor (poświęcony hotelarstwu) czy usług lekarskich (znanylekarz.pl). Pytanie jest zatem następujące: czy można stosować porów-

[144] Tamże, s. 51.

[145] A. Trzoss, *Przyczynek do badań nad metodami historii cyfrowej w świetle debaty przed EU Referendum na profilach portalu Facebook Davida Camerona i Nigela Farage'a*, „Przegląd Archiwalno-Historyczny" 2018, t. 5, s. 203–225.

[146] M. Troszyński, *Ukraina i Ukraińcy w polskim dyskursie internetowym*…, s. 118.

[147] Tenże, *Ukraina i Ukraińcy w polskich mediach społecznościowych*…, s. 54–55.

nanie pomiędzy określeniem stosunku do usługi hotelowej czy medycznej z postrzeganiem przedstawicieli innego narodu? Również automatycznie zastosowany klasyfikator (tym razem wybrano pięć kategorii tematycznych) po czterokrotnej walidacji uzyskał średnio 56% odpowiedzi. Nie wspomniano także, co z wielotematycznością wypowiedzi, tj. czy jeśli dany tekst zawierał w sobie kilka tematów, to czy był przypisywany do jednego z nich (np. uśrednione wektory wskazywały na dominację jednego tematu) czy też przypisywano jeden wpis do kilku grup tematycznych. Powyższe passusy nie mają na celu krytyki czy też walidacji założeń metodologicznych autora, a jedynie wskazanie na delikatność w stosowaniu narzędzi badań ilościowych, gdzie kontekst i kwestia przystawania metod do materiału badawczego (np. proces jego powstawania, ewolucji, zawiłości) jest równie ważny jak możliwości, które dostarcza automatyzacja badań.

Ujęcie mieszane politologiczno-socjologiczne. Czy *echo chamber* naprawdę istnieje?

Paweł Matuszewski jest socjologiem oraz praktykiem *data science*. W swoich badaniach analizuje zjawiska zachodzące na przecięciu się mediów społecznościowych (głównie Facebook oraz Twitter) oraz partii politycznych i mediów tradycyjnych. Łączy on analizę językową z zachowaniem użytkowników względem badanych podmiotów. W pracy poświęconej weryfikacji hipotezy o *echo chambers* (tj. użytkownicy poszczególnych stron fanowskich o charakterze politycznym preferują powoływanie się na media bliskie im ideologicznie za pomocną tzw. odsyłaczy np. linków) w latach 2015–2017[148]. Przeanalizowanych zostało 2,3 mln komentarzy zawierających 140 tysięcy odsyłaczy (w szczegółowej analizie wzięto pod uwagę tylko te media informacyjne, do których użytkownicy odwoływali się co najmniej 10 razy, a zatem 33 tysiące odsyłaczy). Wartością dodatnią w opracowanej przez autora metodologii badania jest wzięcie pod uwagę kontekstu, w jakim został osadzony odsyłacz (dzięki algorytmowi iSAX[149]), a także sentyment. Matuszewski wziął pod uwagę 10 słów przed i 10 po zamieszczonym odsyłaczu, co zostało poparte analizą eksploracyjną i innymi badaniami. Podobnie, jak w przypadku badań Troszyńskiego dokonano półautomatycznego kodowania, jednakże na większej próbie treningowej (zakodowano ręcznie 5000 wpisów) i uzyskano lepsze wyniki (błąd standardowy na poziomie jednego procenta). Autor zaznacza, iż ze względów technicznych nie zostały wzięte pod uwagę odsyłacze prowadzące do portalu Youtube (43 tys.), gdyż ich określenie przynależności wymagałoby ręcznej weryfikacji, co byłoby nazbyt czasochłonne w stosunku do korzyści.

W kwestii uzyskanych wyników przedstawiono różnicę między średnimi polubień pod komentarzami na fanpage'ach partii politycznych, zawierających odnośniki do mediów z podziałem na media lewicowe, prawicowe i nietożsamościowe.

[148] P. Matuszewski, *Wykorzystanie mediów informacyjnych w dyskusjach politycznych na Facebooku*, „Studia Medioznawcze" 2018, nr 1, s. 27–43.

[149] Tamże, s. 30.

W ten sposób autor dokonał falsyfikacji hipotezy, iż użytkownicy powołujący się na media niezgodne z ideologiczną linią partii spotykają się z negatywnym odbiorem[150]. W otrzymanych rezultatach zauważono między innymi, iż na stronach powiązanych z PiS najbardziej lubiane komentarze z odsyłaczami kierowały do mediów o charakterze lewicowym. Analiza kontekstu wykazała, iż w prawie 70% przypadków sentyment towarzyszący odsyłaczowi miał charakter neutralny, a różnica między pozytywnym i negatywnym wynosiła 17% na korzyść pozytywnego.

Matuszewski doszedł do następujących wniosków: nie jest prawdą, iż użytkownicy stron fanowskich partii politycznych częściej powołują się na media związane z nimi światopoglądowo; użytkownicy także nie „lubią" bardziej tychże; wreszcie, sentyment negatywny wobec takowych był stosunkowo rzadką przypadłością (7%). Badanie to poprowadziło do dwóch ciekawych spostrzeżeń. Po pierwsze, zjawisko *echo chambers* (komory pogłosowej) jest w przypadku stron fanowskich polskich partii politycznych zjawiskiem raczej nieistniejącym. Przeciętny użytkownik zderza się jednak z pluralizmem poglądów i argumentów. Po drugie, brak występowania powyższego zjawiska pogłosu nie musi przeczyć koncepcji polaryzacji społecznej. Matuszewski zauważa, iż takowe dyskusje między zwolennikami przeciwnych partii mogą dodatkowo utwierdzać w przekonaniu wyborców danych opcji politycznych.

W drugim badaniu Matuszewski rozwija wątek zaangażowania użytkowników w treści publikowane na Facebooku[151]. Badaniu poddano 1,4 mln użytkowników, którzy „polubili" 30 mln razy analizowane treści. Rezultatem przeprowadzonej analizy jest spostrzeżenie, iż użytkownicy angażują się sporadycznie w treści umieszczane pod profilami partii politycznych, polityków i mediów. Co więcej, mediana zaangażowania w różne profile wynosiła 5, a zatem pojawia się hipoteza, iż użytkownicy wybierają wąską grupę, jasno określonych światopoglądowo stron, których treści chłoną. W porównaniu z poprzednim badaniem oznacza to, iż co prawda na wybranych profilach użytkownicy spotykają się w komentarzach z opiniami „drugiej strony", ale większość z nich sama chłonie skondensowane i jednostronnie prezentowane informacje.

Innym podejściem do analizy treści jest badanie powiązań między odbiorcami treści, a ich kreatorami[152]. W tym celu wykorzystano (stosowaną z sukcesem w analizie samej treści, o czym dalej) analizę powiązań sieciowych na portalu Twitter (*social network analysis*[153]), gdzie punktami węzłowymi były istotne (top 1% pod względem śledzących użytkowników) profile osób związanych z polityką i mediami. Dokonano porównania między przykładem polskim i węgierskim we wrześniu 2018 roku. W przeciwieństwie do badań nad Facebookiem wykazano

[150] Tamże, s. 36.

[151] P. Matuszewski, *Selective Exposure on Polish Political and News Media Facebook Pages*, „Polish Sociological Review" 2019, nr 1, s. 177–197.

[152] P. Matuszewski, G. Szabó, *Are Echo Chambers Based on Partisanship? Twitter and Political Polarity in Poland and Hungary*, „Social Media + Society" 2019, t. 5, nr 2.

[153] *The SAGE Handbook of Social Network Analysis*, red. J. Scott, P.J. Carrington, SAGE, 2011.

tu, iż podziały są o wiele bardziej wyraźne i pokrywają się z podziałem społecznym. Zatem, nie tylko analiza sentymentu czy zaangażowania w treści, ale także sieć powiązań między nimi pozwala lepiej poznać konteksty funkcjonowania treści politycznych w obszarze mediów społecznościowych. Matuszewski wskazuje przy tym (w przeciwieństwie do Troszyńskiego), iż istotnym dla zróżnicowania wyników jest sama mechanika analizowanych generatorów treści. Inne wyniki dla Twittera i Facebooka wiążą się z ich indywidualną specyfiką. Mówiąc dalej, można zastanowić się czy medium, takie jak YouTube, forum internetowe czy blog (które badał Troszyński) również nie charakteryzują się różnicami w stosunku do dwóch wcześniej wspomnianych mediów. To, jak specyfika danego generatora wpływa na powstawanie i funkcjonowanie treści jest co prawda kwestią bardziej medioznawczą niż językoznawczą, niemniej jednak jedno bez drugiego pozostawia badacza bez kontekstu, co może doprowadzić do mylnych wniosków.

Ujęcie literaturoznawcze. Czy stylometrią można rozwiązać „stare" problemy badawcze?

Ostatnim przykładem wykorzystania metod analizy języka naturalnego są prace Macieja Edera[154]. Można by określić ten sposób zastosowania metod jako najbliższy, w swojej wymowie, tradycyjnym badaniom językoznawczym, w tym przypadku w zakresie ustalania autorstwa tekstów. Eder stosuje metodę badań stylometrycznych, która była już znana przed epoką komputeryzacji, jednakże cyfryzacja tej metody badawczej znacząco rozwinęła ten kierunek badań. Autor zwraca uwagę na to, iż badania stylometryczne (w zakresie przypisywania autorstwa tekstom obecnie anonimowym) pozwalają dostrzec niedostrzegalne okiem właściwości i schematy[155].

Maciej Eder wskazuje na kilka problemów i pułapek stojących za stosowaniem metod „nietradycyjnej atrybucji autorstwa"[156]. Dlaczego jednak stylometria jest tak ciekawym zbiorem metod i jakie oferuje możliwości? Przede wszystkim stylometria skupia się na wyłanianiu statystycznych zależności w stosowanym języku. W założeniu tym zwraca się uwagę na te elementy, których autor używa nieświadomie, a na co mają wpływ jego otoczenie kulturowe, wykształcenie, cechy charakteru itd. Przykładem takich elementów są wyrazy synsemantyczne (rodzajniki, spójniki, przyimki, partykuły)[157]. Jednym z kluczowych pojęć w stylometrii jest MFWs, czyli najczęściej występujące słowa (*most frequent words*) w danym tekście. Słowa te są później poddawane wektoryzacji, czyli zamianie w reprezen-

[154] M. Eder, *Mind your corpus: systematic errors in authorship attribution*, „Literary and Linguistic Computing", 2013, t. 28, nr 4, s. 603–614; M. Eder, *Metody ścisłe…*; M. Eder, *In Search of the Author of Chronica Polonorum Ascribed to Gallus Anonymus: A Stylometric Reconnaissance*, „Acta Poloniae Historica" 2015, t. 112, s. 5–23.

[155] Tenże, *In Search of the Author*…, s. 11.

[156] Tenże, *Mind your corpus*…, s. 603–604.

[157] Tenże, *Metody ścisłe…*, s. 93.

tację liczbową, tak by policzyć odległość (bliskość znaczeniową) między danymi wyrazami. Analogicznie można postąpić z całym tekstem, gdzie porównując dwa teksty bada się różnice (uśrednione) częstotliwości występowania danych zwrotów. W omawianej metodzie, Eder zwraca uwagę, iż nie porównuje się dwóch anonimowych tekstów między sobą – tekst nieznanego autorstwa odnosi się do korpusu tekstów, które są przypisane do twórców[158].

Jak już zostało wspomniane, Eder zwraca wyjątkową uwagę na ostrożność w stosowaniu komputerowych metod statystycznych w analizie języka naturalnego. Punktuje on kilka najważniejszych błędów, które mogą się pojawić podczas pracy z tekstami i jak wpływają one na efektywność stosowanej metody atrybucji autorstwa[159].

W pierwszym przypadku pod rozważania wzięto problem błędnie zapisanych znaków. Spowodowane to może być różnymi czynnikami, od złej transkrypcji czy odczytania cyfrowego (OCR). Poziom uszkodzeń wprowadza tzw. szumy czyli taki rodzaj informacji, który utrudnia rozpoznanie „właściwej" warstwy treściowej. W eksperymencie dokonano 100 iteracji, w których z każdą kolejną poszczególne znaki miały odpowiedni procent szans (np. przy iteracji 15 było to 15%) na zamianę w inną losową literę. Badanie przeprowadzone dla tekstów angielskich, niemieckich, polskich i łacińskich miało za zadanie zbadać efektywność metody atrybucji autorstwa względem uszkodzenia tekstu. Biorąc pod uwagę MFWs na poziomie 500 słów (500 najczęściej występujących wyrazów) zaobserwowano, iż algorytm poprawnie przypisuje autorstwo mimo znacznych uszkodzeń. W niektórych przypadkach nawet 20% szumu nie powodowało większych problemów[160]. Zwrócono jednak uwagę na zależność pomiędzy długością słów a spadkiem efektywności – tu problem występował głównie dla tekstów niemieckojęzycznych.

Analogicznym przypadkiem było wprowadzenie szumu w postaci dodatkowych słów w tekście. Ich pochodzenie mogłoby wynikać z poprawek edytorskich lub późniejszych komentarzy. W wyniku eksperymentu zauważono pewną paralelność – w tym przypadku z kolei, krótsze słowa były bardziej podatne na spadek efektywności, zaś dłuższe bardziej odporne, gdyż mają mniejsze przełożenie na przypisywanie autorstwa (występują rzadziej).

Najbardziej intrygujący był jednak eksperyment trzeci, w którym Eder zwrócił uwagę na bardziej tradycyjną stronę językoznawstwa, jaką jest wpływ tradycji literackich na teksty[161]. Dokonano symulacji wpływu inspiracji (plagiaty, imitacje itd.) na przypisywanie autorstwa. Założono tu, iż błędnym byłoby stwierdzenie, iż każdy autor jest niepowtarzalny, a jego dzieła oryginalne tak, jakby nie istniały żadne inne. W 100 iteracjach wprowadzono (analogicznie do eksperymentu 1 i 2) szum w postaci zapożyczeń fraz z innych tekstów z korpusu, zwiększając tym powiązania między anonimowym tekstem a dziełami, do których go odnoszono. W przypadku języków nowożytnych zaobserwowano znaczący spadek skutecz-

158 Tamże, s. 95.
159 Tenże, *Mind your corpus*…
160 Tamże, s. 605–606.
161 Tamże, s. 607.

ności metody przypisywania autorstwa. Wyjątkiem okazały się teksty łacińskie, gdzie paradoksalnie mimo nawet 40% szumu atrybucja autorstwa była skuteczna. Eder konkluduje, iż „unikalność" autorów łacińskich opiera się (w tym eksperymencie) na wykorzystaniu cytatów z innych dzieł. Innymi słowy, o unikalności autora (a zatem przypisaniu mu tekstu) świadczył nie „specyficzny" dla niego styl, a jego zdolności literackie w zakresie znajomości tradycji piśmienniczych.

Jak już zauważono, poprawnie stosowane metody z uwzględnieniem kontekstów i świadomość ich ograniczeń, pozwalają na skuteczniejsze ich stosowanie w badaniach. Maciej Eder poświęcił temu zagadnieniu osobny artykuł przestrzegając humanistykę przed ślepą wiarą w obiektywizm i możliwości, jakie daje stosowanie metod nauk ścisłych[162]. Zasadniczą myślą, wokół której Eder konstruuje swoją narrację, jest problem założeń obiektywizmu metod nauk ścisłych tj. powtarzalność i weryfikowalność badań.

Po pierwsze, dobieranie wartości przez badacza odbywa się arbitralnie w zależności od materiału badawczego. Przy doborze zakresu MFWs mogą przynieść różne wyniki np. dla wartości 500 czy 1000 najczęściej występujących słów[163]. Analogicznie używanie innego algorytmu, filtracji danych czy wrażliwość na wcześniej wspomniane szumy również wpływa na obiektywizm badania. Niestabilność wyników jest jednym z czołowych problemów, z którym różnie sobie radzono. Stosuje się metody konsensusowe, czyli uśrednienie wyników. Eder przyrównuje tę optykę do postrzegania katedry Notre-Dame w Rouen, a mianowicie, że dopiero wiele różnych perspektyw pozwoli uzyskać lepszy obraz sytuacji, co jednak nie powoduje absolutnej obiektywności[164]. Badania stylometryczne są zatem uzupełnione o wcześniej wspomniane badania sieci (*social networking analysis*), gdzie punktami na przestrzeni są teksty, a połączenia między nimi to podobieństwa między różnymi parametrami np. MFWs. Przeprowadza się kilka takowych eksperymentów, a następnie nakłada na siebie[165]. Im „grubsze" i krótsze połączenie, tym silniejsze podobieństwo między tekstami. Eder konkluduje, iż mimo uzyskania pewnych rozwiązań dla problemu niestabilności i wizualizacji wyników to nadal kwestią arbitralną pozostaje dobór i zakres parametrów przez poszczególnych badaczy. Eder twierdzi, iż: „Obiektywizm w sensie ścisłym, wolno sądzić, pozostanie na zawsze poza zasięgiem badań literaturoznawczych – ze względu na charakter badanego materiału, czyli artystycznie ukształtowanej kreacji literackiej"[166].

Wpisują się w to stwierdzenie poprzednie uwagi zawarte wobec innych, nakreślonych w tym tekście, obszarów badawczych. Samo stosowanie metody, choćby skutecznie działającej w naukach ścisłych, wobec materiału z obszaru humanistyki, niewiele daje, a może prowadzić do mylnych interpretacji. Eder odradza m.in. stosowanie (spotykanej praktycznie w każdym badaniu NLP) lematyzacji wobec

[162] Tenże, *Metody ścisłe*...
[163] Tamże, s. 99.
[164] Tamże, s. 101.
[165] Tamże, s. 101–104.
[166] Tamże, s. 104.

tekstów pisanych po łacinie[167]. Argumentuje to faktem, iż w przypadku łaciny formy słów jak „być" pełnią rolę cech charakterystycznych stylu danego pisarza.

Powyższe rozważania służyły naszkicowaniu tła dla badania, które przeprowadził Maciej Eder, a które zakładało weryfikację hipotezy Tomasza Jasińskiego, jakoby autor *Chronica Polonorum* był „mnichem z Lido"[168]. Eder stosując inną metodę niż Jasiński. Założył, iż jeśli obie dadzą zbliżony wynik, to hipoteza „mnicha z Lido" ma spore prawdopodobieństwo na bycie poprawną[169].

Stosując metody stylometryczne, Eder początkowo przyrównał tekst Kroniki do innego tekstu autorstwa mnicha z Lido. Mimo początkowych różnic między oboma dziełami postanowiono poszerzyć korpus i zestawić oba ze 159 tekstami łacińskimi, aby sprawdzić czy te różnice utrzymują się w stosunku do większego zbioru literackiego, czy też korelują z innymi pracami. Połączona stylometria i analiza powiązań sieciowych dała zaskakująco pozytywny wynik, potwierdzając hipotezę Jasińskiego. Co więcej, analogiczna siła powiązań między tekstami była zaobserwowana jedynie w przypadku tekstów autorstwa tej samej osoby (Cycerona czy Witruwiusza)[170].

Podsumowanie. Czy metody analizy języka naturalnego mogą być uniwersalne?

Pytanie postawione w tytule zakończenia jest nieco przewrotne. Nie sposób w krótkim artykule, czy nawet sporych rozmiarów monografii wskazać na wszystkie problemy badawcze, do których można zastosować metody NLP. Jednakże tym, na co chciałbym zwrócić uwagę, to kolejne możliwości, jakie otwierają się dzięki stosowaniu omawianych metod. W ramach podsumowania chciałbym wskazać dwa niedawno opublikowane przykłady.

W pierwszym, Marcin Wilkowski na swoim blogu prezentuje możliwości, jakie daje analizowanie nagłówków tekstów na portalach zajmujących się popularyzowaniem historii[171]. Zestawiając prawie trzy tysiące tekstów z dwóch najpopularniejszych portali, Wilkowski przeanalizował częstotliwość pojawiania się kluczowych terminów w nagłówkach. Hipotezą postawioną w badaniu była relacja między *clickbaitem* a historycznymi odniesieniami. Okazało się, iż poza stosowaniem chwytów retorycznych (stosowanie pytań w tytule, popularność zwrotu „naprawdę" w tekstach) istotne miejsce mają kontrowersyjne postacie historyczne, pośród których czołowe miejsce zajął Adolf Hitler. Po pierwsze, Wilkowski zwraca uwagę na kwestię recepcji postaci historycznych w literaturze popularnonaukowej w Internecie. Po drugie, na zainteresowania oddolne czytelników, którzy

[167] Tenże, *In search of the author…*, s. 12.
[168] Tamże.
[169] Tamże, s. 7.
[170] Tamże, s. 22.
[171] M. Wilkowski, *Histmag vs Ciekawostki Historyczne: clickbait i pisanie o przeszłości*, opublikowano: 04.07.2016, https://wilkowski.org/notka/1267 [dostęp: 28.05.2020].

chętniej sięgają po tytuły odwołujące się do kwestii kontrowersyjnych, przemocy czy ludzkich dramatów[172].

Druga praca to artykuł Romana Deikslera poświęcona stosowaniu analizy powiązań sieciowych w badaniach nad dziełami Józefa Flawiusza[173]. Deiksler wskazuje, iż dzięki wspomnianej metodzie, udało się wskazać na zależności między znaczeniem poszczególnych miast w trakcie powstania żydowskiego w Galilei (I w. n.e.) a aktorami wydarzenia. Analiza ta wskazała na szczególne miejsce Gamli, jako istotnej fortecy, oraz Sepphoris. Stosowanie analizy powiązań sieciowych może dotyczyć zatem nie tylko tekstów czy słów, ale i postaci czy miejsc.

Konkludując, chciałbym zwrócić raz jeszcze uwagę na najistotniejsze punkty, które zostały pokrótce zarysowane w moich rozważaniach. Po pierwsze, problem gromadzenia i selekcjonowania materiału źródłowego tak, aby nie tylko możliwe było stosowanie metod, ale i by wnioski uzyskane za ich pomocą korespondowały z opisywaną rzeczywistością. Po drugie, zwrócenie uwagi na kontekst powstawania i funkcjonowania omawianych zjawisk i materiału badawczego. Różne metody i wnioski mogą odnosić się do różnie funkcjonujących mediów. Po trzecie, świadomość ograniczeń oraz pułapek czyhających na niewprawnych badawczy oraz zapatrzenie w obiektywizm i efektywność stosowanych metod w naukach ścisłych. Należy pamiętać, jak zaznaczył Maciej Eder, iż metody te pierwotnie zostały opracowane z myślą o innych typach materiału badawczego, stąd dopasowanie metod do materiału źródłowego powinno się mieć na uwadze. Wreszcie, możliwości, jakie idą za mariażem metod nauk ścisłych a humanistyką i naukami społecznymi, wydają się być obiecujące. Transdyscyplinarność i interdyscyplinarność jest nie tylko modnym obecnie paradygmatem, ale jak pokazano w przypadku Galla Anonima, może pomóc powrócić do starych problemów badawczych i wspomóc nad nimi prace. Jak zaznaczono we wstępie, niniejszy artykuł miał za zadanie przybliżyć polskiemu czytelnikowi polskie wątki badań stosujących przetwarzanie języka naturalnego. Mimo pobieżnego przedstawienia aktualnego stanu literatury mam nadzieję, iż udało się wskazać na pożytek lektury tekstów w tym obszarze.

Bibliografia

Babik W., *Język naturalny w wyszukiwaniu informacji i problemy jego przetwarzania*, „Zagadnienia Informacji Naukowej" 2013, nr 1, s. 37–47.

Bartusiak R., Augustyniak Ł., Kajdanowicz T., Kazienko P., Piasecki M., *WordNet2Vec: Corpora agnostic word vectorization metod*, „Neurocomputing" 2019, nr 326–327, s. 141–150.

[172] Szerzej zob. W. Werner, D. Gralik, A. Trzoss, *Media społecznościowe a funkcjonowanie wiedzy historycznej w Polsce. Raport z badań*, „Przegląd Archiwalno-Historyczny" 2019, t. 6, s. 211–235; W. Werner, A. Trzoss, D. Gralik, *Historia i YouTube. Narracja historyczna w dobie Web 2.0*, „Nauka" 2020, nr 3, s. 119–140.

[173] R. Deiksler, *Social Network Analysis In The Study of The Works Of Josephus. The Case Study Of Galilee During The First Jewish Revolt*, „Folia Praehistorica Posnaniensia" 2019, t. 24, s. 35–46.

Deiksler R., *Social Network Analysis In The Study of The Works Of Josephus. The Case Study Of Galilee During The First Jewish Revolt*, „Folia Praehistorica Posnaniensia" 2019, t. 24, s. 35–46.

Eder M., *In Search of the Author of Chronica Polonorum Ascribed to Gallus Anonymus: A Stylometric Reconnaissance*, „Acta Poloniae Historica" 2015, t. 112, s. 5–23.

Eder M., *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii*, „Teksty Drugie" 2014, nr 2, s. 90–105.

Eder M., *Mind your corpus: systematic errors in authorship attribution*, „Literary and Linguistic Computing" 2013, t. 28, nr 4, s. 603–614.

Malak P., *Rozwój badań nad przetwarzaniem języka naturalnego*, „Zagadnienia Informacji Naukowej" 2010, nr 2, s. 21–30.

Matuszewski P., *Selective Exposure on Polish Political and News Media Facebook Pages*, „Polish Sociological Review" 2019, nr 1, s. 177–197.

Matuszewski P., Szabó G., *Are Echo Chambers Based on Partisanship? Twitter and Political Polarity in Poland and Hungary*, „Social Media + Society" 2019, t. 5, nr 2.

Matuszewski P., *Wykorzystanie mediów informacyjnych w dyskusjach politycznych na Facebooku*, „Studia Medioznawcze" 2018, nr 1, s. 27–43.

The SAGE Handbook of Social Network Analysis, red. J. Scott, P.J. Carrington, SAGE, 2011.

Troszyński M., *Analiza treści witryn internetowych z wykorzystaniem automatycznego kodowania* [w:] *Metody badań online*, red. P. Siuda, Gdańsk 2016, s. 83–103.

Troszyński M., *Ukraina i Ukraińcy w polskim dyskursie internetowym. Analiza jakościowo-ilościowa tekstów zamieszczanych w mediach społecznościowych* [w:] *Raport. Mniejszość ukraińska i migranci z Ukrainy w Polsce. Analiza dyskursu*, red. P. Tyma, Warszawa 2018, s. 103–185.

Troszyński M., *Ukraina i Ukraińcy w polskich mediach społecznościowych* [w:] *Raport 2. Mniejszość ukraińska i migranci z Ukrainy w Polsce. Analiza dyskursu*, red. P. Tyma, Warszawa 2019, s. 49–76.

Trzoss A., *Przyczynek do badań nad metodami historii cyfrowej w świetle debaty przed EU Referendum na profilach portalu Facebook Davida Camerona i Nigela Farage'a*, „Przegląd Archiwalno-Historyczny" 2018, t. 5, s. 203–225.

Werner W., Gralik D., Trzoss A., *Media społecznościowe a funkcjonowanie wiedzy historycznej w Polsce. Raport z badań*, „Przegląd Archiwalno-Historyczny" 2019, t. 6, s. 211–235.

Werner W., Trzoss A., Gralik D., *Historia i YouTube. Narracja historyczna w dobie Web 2.0*, „Nauka" 2020, nr 3, s. 119–140.

Wilkowski M., *Histmag vs Ciekawostki Historyczne: clickbait i pisanie o przeszłości*, opublikowano: 04.07.2016, https://wilkowski.org/notka/1267 [dostęp: 28.05.2020].