2021 PRACE JĘZYKOZNAWCZE XXIII/4
ISSN 1509-5304 DOI 10.31648/pj.7050 295–302

Michał Szczyszek
Uniwersytet im. Adama Mickiewicza w Poznaniu
ORCID: https://orcid.org/0000-0002-0253-7296
e-mail: szczysze@amu.edu.pl

Filip Graliński: *Against the Arrow of Time. Theory and Practice of Mining Massive Corpora of Polish Historical Texts for Linguistic and Historical Research*. Wydawnictwo Naukowe UAM. Poznań 2019, ss. 315

Przedmiotem niniejszej recenzji jest książka Filipa Gralińskiego *Against the Arrow of Time. Theory and Practice of Mining Massive Corpora of Polish Historical Texts for Linguistic and Historical Research*. Sytuuje się ona na pograniczu dwóch obszarów badawczych: 1) informatyki, zwłaszcza: przetwarzania języka naturalnego oraz 2) lingwistyki, zwłaszcza korpusologii (diachronicznej) – w tym: lingwochronologii.

Praca jest bardzo cennym opracowaniem pokazującym problematykę, metodologię i procedury związane z budowaniem cyfrowych narzędzi służących humanistyce, procesem weryfikacji zarówno danych wchodzących (tzw. inputowych), procesu ich analizy za pomocą narzędzia cyfrowego oraz danych uzyskanych w wyniku działania narzędzia cyfrowego (tzw. outputowych), które składają się na nieustannie rozrastający się korpus polszczyzny. Autor swoje rozważania oparł na doświadczeniu i obserwacjach związanych ze stworzonym przez siebie narzędziem – systemem „Odkrywka". Narzędzie to gromadzi, przetwarza na podstawie stosownych algorytmów obszerne dane językowe z języka polskiego. Ich obszerność zobrazować można informacjami o zasięgu chronologicznym: w korpusie znajdują się dane od początku XIX w. do dziś (por. s. 42 monografii); pojawiają się także dane z okresu wcześniejszego. Dodać należy, jak sam autor deklaruje, że ten korpus historyczny będzie się powiększał wraz z rozwojem bibliotek cyfrowych, w których będą się pojawiały zdygitalizowane teksty z doby nowopolskiej, średniopolskiej i wcześniejszych okresów rozwojowych polszczyzny.

Dane liczbowe obrazujące wielkość korpusu „Odkrywka", to – podajmy za autorem (a parafrazuję jego angielskojęzyczne sformułowania w moim własnym tłumaczeniu; robię tak za każdym razem, gdy w recenzji przytaczam lub parafrazuję zapisy w języku angielskim), który pisze, że całkowita ilość tekstu przetworzonego dla narzędzia Odkrywka wynosi 96 368 642 437 znaków, tj. 15 137 368 095 słów, a słowo definiuje się tu jako ciągłą sekwencję liter i cyfr (s. 63). Ponad 15 mld wyrazów tekstowych (realizacji leksemowych) czyni z tego korpusu największy zbiór wyrazów dla języka polskiego i jeden z największych ze znanych i dostępnych korpusów światowych (może równać się z ogromnymi korpusami języka angielskiego, np.: „NOW Corpus" („News on the Web", który – jak czytamy na jego stronie (https://www.english-corpora.org/now/) – zawiera 10 mld słów ekscerpowanych z internetowych wydań gazet i czasopism od 2010 r. do chwili obecnej (ostatni dzień to 2020-05-20)) czy korpusem „iWeb corpus" (https://www.english-corpora.org/iweb/, który zawiera 14 mld słów z 22 mln stron internetowych).

Należy oczywiście zauważyć różnicę jakościową zachodzącą między korpusem stworzonym przez Filipa Gralińskiego a przywołanymi korpusami języka angielskiego. Wynik porównania wypada na korzyść korpusu „Odkrywka". Obejmuje on swoim zasięgiem chronologicznym kilkaset lat rozwoju polszczyzny, a korpusy angielskie – około jednej dekady(!). Korpusy angielskie bazują na danych tylko internetowych (czyli tzw. digital-born), natomiast korpus „Odkrywka" gromadzi dane pochodzące z tekstów: drukowanych, a następnie zdygitalizowanych i poddanych procedurze OCR, ręcznie przepisywanych oraz pochodzących wprost z internetu. Z danymi językowymi pochodzącymi z procesu digitalizowania druków wiążą się kolejne niemałe problemy techniczne, metodologiczne, proceduralne, które w recenzowanej książce zostały szczegółowo rozpatrzone.

Dla porządku recenzenckiego dodać należy i to, że w korpusie zgromadzone są dane językowe pochodzące z różnych odmian polszczyzny pisanej – zarówno z prasy, z literatury pięknej, z internetu, jak i ze stenogramów sejmowych, listów, dokumentów umieszczanych w postaci zdygitalizowanej w polskich bibliotekach cyfrowych, druków ulotnych (por. PART I. TEXTUAL MASS, s. 19–31).

Zatem już teraz można powiedzieć, że narzędzie korpusowe Gralińskiego należy ocenić jako osiągnięcie naukowe klasy światowej w kategorii prac z zakresu lingwistyki komputerowej. Daje ono ogromne możliwości badawcze w zakresie np. lingwochronologii, badań fleksyjnych (np. ustalanie wariantów odmiany, orzekanie w sprawie dominacji innowacji rozszerzających i/lub regulujących), leksykalno-semantycznych (np. określanie przemian znaczeniowych, ustalanie dominant semantycznych jednostek leksykalnych),

słowotwórczych (np. obserwacja rozprzestrzeniania się w tekstach w różnych okresach poszczególnych technik słowotwórczych), stylistycznych (np. w zakresie „śledzenia" sformułowań kolokwialnych, potocznych przenikających do tekstów i/lub gatunków wypowiedzi, które mają naturę i właściwości bardziej oficjalne), czy wyszukiwanie motywów literackich lub etnologicznych (np. związanych z legendami miejskimi).

Recenzowana książka, opisująca zarówno autorski korpus (autorskie narzędzie) „Odkrywka", jak i możliwe procedury badawcze wykorzystujące ten korpus, to z całą pewnością wyważona, logicznie spójna, precyzyjna i przejrzysta pod względem kompozycyjnym monografia składającą się ze *Wstępu* (s. 11–19), jedenastu rozdziałów pogrupowanych w cztery części oraz składników kończących i dopełniających monografię: *List of excerpts, List of figures*, *List of tables*, *Indeksu* oraz obszernej *Bibliografii* (te końcowe składniki monografii zapisane są na s. 287–315). Części te, to – wraz z rozdziałam: PART I. *TEXTUAL MASS* (s. 19–101): Chapter 1. *What is out there?*, Chapter 2. *Metadata*, Chapter 3. *Texts*; PART II. *(RE)SEARCHING* (s. 103–150): Chapter 4. *Searching for words*, Chapter 5. *From search into research*; PART III. *MODELLING* (s. 151–170): Chapter 6. *Temporal language models*, Chapter 7. *Temporal text classification*, Chapter 8. *Word embeddings for diachrony*; PART IV. *APPLICATIONS* (s. 225–285): Chapter 9. *Lexical ephemera*, Chapter 10. *Traps of culturomics*, Chapter 11. *Folkloristics 2.0.*

We *Wstępie* autor wprowadza czytelnika w problematykę monografii, precyzując przedmiot swoich dociekań, przedstawiając założenia teoretyczno-metodologiczne, zgodnie z którymi przeprowadzał ekscerpcję i gromadzenie materiału, analizy oraz wnioskowanie. Przedstawił też główne cele swoich badań.

Następnie w części: PART I. *TEXTUAL MASS* przedyskutowane są i dogłębnie omówione zagadnienia związane z pozyskiwaniem tekstów źródłowych, normalizacją ich metadanych i procesami rozpoznawania tekstów przez cyfrowe narzędzie (np. OCR) na potrzeby przygotowania pełnotekstowego wyszukiwania. Daje to czytelnikowi dobrą perspektywę poznawczą w zakresie cyfrowego gromadzenia danych językowych i ich przetwarzania. Graliński bardzo precyzyjnie wskazał źródła tekstów – pokazał, że wyzyskał materiał zgromadzony w postaci zdygitalizowanej w większości polskich bibliotek cyfrowych, z uwzględnieniem tych najobszerniejszych i najbardziej rozpowszechnionych wśród użytkowników (np. Wielkopolska BC), jak i tych mniej znanych (np. Cyfrowa Biblioteka Druków Ulotnych). Uzyskał w ten sposób nie tylko obszerną bazę tekstową, ale także – co niezwykle istotne – bardzo zróżnicowaną. Następnie autor opisał opracowaną przez siebie procedurę normalizacji metadanych m.in. w zakresie datowania tekstów

(np. przy niepewnej datacji), tytułów tekstów, rodzajów publikacji, co jak wiadomo w pracach korpusologicznych jest niezwykle istotne i wymaga dużej akrybii. Dzięki tej procedurze dane zgormadzone w korpusie „Odkrywka" (np. ich datowanie) nie są narażone na swoiste „podróże w czasie". Ostatnim punktem części pierwszej monografii jest opis działań zmierzających do uczynienia zdygitalizowanych druków czytelnymi dla narzędzi cyfrowych, czyli opis opracowanej samodzielnie i zastosowanej przez autora informatycznej metody związanej z NLP: począwszy od „oczyszczenia" wyników pracy programu OCR, a skończywszy na przygotowaniu wyszukiwarki pełnotekstowej. Wszystkie te działania, opracowane procedury należy uznać za pionierskie na gruncie cyfrowego przetwarzania tekstów polskich; niewykluczone, że okazałoby się, że w skali światowej niewiele jest podobnych rozwiązań, zwłaszcza o tak wysokim poziomie skuteczności (por. np. zapisy na s. 81 monografii).

W dalszej części: PART II. *(RE)SEARCHING* autor przedstawił metodologię pozyskiwania danych językowych z opracowanego przez siebie narzędzia „Odkrywka". Metodologię tę w istocie oddaje tytuł rozdziału piątego (należącego do tej części): *From search into research* („Od wyszukiwania do badania"). Autor pokazał bowiem w całej części drugiej swojej monografii zarówno procedurę pozyskiwania (wyszukiwania) jednostek językowych zgromadzonych w korpusie „Odkrywka", możliwości uzyskiwania danych statystycznych, dających się pozyskać z tego ogromnego korpusu, jak i „zdolność" tego narzędzia do tworzenia całościowego *dossier* danej jednostki językowej. Innymi słowy – pokazał, jak i jakie informacje można pozyskać oraz jak je można/należy interpretować, aby uzyskać właściwy, tj. prawdziwy obraz języka (danej jednostki językowej) i nie ulec swoistej pokusie nadinterpretacji w opisanych zakresach. W tej części monografii autor przeprowadza swoisty instruktaż opisujący wszystkie wskazane tu aspekty pracy badawczej z „Odkrywką", unaoczniając i przestrzegając przed owymi pomyłkowymi interpretacjami.

Kolejna część PART III. *MODELLING* wydaje się niezwykle ważna z punktu widzenia połączenia kompetencji informatycznych z lingwistycznymi, zwłaszcza dotyczącymi badań historycznojęzykowych. Potrzeba dostosowania modelu informatycznego do danych historycznych języka polskiego wymusiła na autorze opracowanie maszynowego (informatycznego) modelu ewaluacji danych historycznojęzykowych, a w konsekwencji i proces odwrotny, tj. dostosowanie diachronicznego modelu samego języka do możliwości narzędzia cyfrowego, w tym – do możliwości *uczenia maszynowego* dla tego typu danych. Zadanie niełatwe, wymagające niemałych kompetencji informatycznych i takowej wyobraźni badawczej, zwłaszcza że Graliński nie jest

filologiem, a właśnie informatykiem. Niekoniecznie więc miał możliwość wcześniejszego wypracowania na własne potrzeby swojej wizji ewolucji języka polskiego i jego stadiów wcześniejszych. Opierając się zatem na dostępnych opracowaniach, wybrał i dostosował do swoich potrzeb te, które z informatycznego punktu widzenia najlepiej wpisywały się w możliwości technologiczne związane z lingwistyką komputerową. Wypracował zatem specjalne rozwiązanie: RetroGapo, które – jak pisze autor – zostało przygotowane i udostępnione na platformie Gonito.net platform (https://gonito. net/challenge/retro-gap) jako właśnie diachroniczny model języka dostosowany do możliwości maszynowych (uczenia maszynowego). Dane te – jak dalej pisze Graliński – są łatwo dostępne jako repozytorium pod adresem git://gonito.net/retro-gap.git i zostały one przygotowane przy użyciu ogólnych ram generowania informatycznych „wyzwań" diachronicznych dla języka polskiego. Dane te – dotyczące procedury dostosowawczej – pobrano z „Odkrywki" (z lat 1814–2013). Korpusy RetroGap zostały podzielone na zestaw treningowy, dwa zestawy rozwojowe (dev-0 i dev-1) oraz zestaw testowy RetroC corpus (por. s. 164). RetroC corpus, to: „[…] to polskojęzyczny korpus diachroniczny przeznaczony do szkolenia i testowania automatycznych systemów datowania" (por. s. 177). I dalej Graliński doprecyzowywał: „Do tej pory opracowałem dwie wersje korpusu: RetroC1 w 2015 r. I drugą – RetroC2 w 2017 r. RetroC2 jest nie tylko większy (jest nadzbiorem RetroC1), ale zawiera również dodatkowe funkcje w zestawie treningowym. Korpus został zaprojektowany z myślą o następujących celach:

- ma być zbiorem polskich tekstów;
- ma być zbiorem wystarczająco dużym, aby umożliwić stosowanie metod statystycznych;
- ma być zbiorem rozległym w czasie – zawierającym nie tylko nowoczesne teksty internetowe, ale także dawne materiały drukowane;
- ma umożliwiać opracowywanie maszynowe także i krótkich tekstów, nie tylko całych książek (opracowywanie maszynowe, tj. np. datowanie całych książek jest znacznie łatwiejsze) (por. s. 177).

To pozwoliło przyjąć „Word2vec model", o którym autor szerzej pisze na s. 188 jako o zadaniu matematyczno-informatycznym. *Notabene* – cała druga część monografii składa się z bardzo szczegółowych rozważań matematyczno-informatycznych w odniesieniu do istoty języka.

Przy takim podejściu badawczym (zawierającym istotne elementy pragmatyki naukowej) Graliński mógł aposteriorycznie wygenerować swoje własne spojrzenie na język (i jego ewolucję), na który patrzy poprzez wymogi (i ograniczenia?) NLP. Pomysł na stworzenie systemu uczącego się, trenowanego na językowym materiale zadanym, dał w efekcie bardzo precyzyjne narzędzie

cyfrowe (właśnie system „Odkrywka") umożliwiające precyzyjne badania w zakresie polskiej leksyki historycznej (i szerzej: w zakresie historii polszczyzny).

W części ostatniej, PART IV. *APPLICATIONS*, Graliński zaprezentował możliwości praktyczne związane z wykorzystaniem stworzonego przez siebie narzędzia. Ta część to zbiór *case studies* związanych z jednej strony z poszukiwaniem efemerycznych struktur językowych (por. Chapter 9. *Lexical ephemera*) – a więc struktur znanych historykom języka jako *hapaks legomenon*, a drugiej strony – folklorem (tu: miejskim), czy – jak sam autor pisze – z folklorystyką 2.0 (odwołując się przy okazji do pojęcia kulturomics; por. Chapter 10. *Traps of culturomics*). Nie wchodząc w dyskusję z dotychczasowymi próbami odnajdywania *hapaks legomena* (np. w słownikach notujących pewien procent „produkcji" językowej danego czasu), można pokusić się o przyjęcie założenia, że metody badawcze Gralińskiego oraz możliwości tkwiące w jego korpusie pozwolą na faktyczne odnalezienie efemerycznych struktur językowych danego okresu rozwojowego polszczyzny. Osiągnięcie to samo w sobie jest bardzo wartościowe, a w monografii pokazane na kilku przykładach. Podobnie rzecz się ma z możliwościami śledzenia wątków folklorystycznych. Autor pokazał na kilku przykładach możliwości praktyczne tkwiące w opracowanym przez niego narzędziu, co z pewnością dać może mocny impuls do rozwoju badań etnologicznych i folklorystycznych w ujęciu 2.0.

Mankamentem monografii może być to, że nie zaprezentowano w niej zbyt dużej liczby analiz ściśle językoznawczych. Przykładowo: autor nie omawia większej liczby jednostek efemerycznych (może ich więcej nie ma? nie wiadomo). W ogóle niewiele miejsca (w odniesieniu do całości monografii) poświęca na zaprezentowanie wyników analiz językoznawczych możliwych do przeprowadzenia z wykorzystaniem „Odkrywki". Nie jest oczywiście tak, że tych analiz nie ma. Ich przykładem są opisy, wykresy i ich interpretacje chronologii wystąpień wyrazu *telewizja* (por. s. 113), a także – cała część czwarta *APPLICATIONS*. Można by w tym miejscu napisać, że oczekiwany byłby jakiś minisłownik czy próbka danych językowych, aby czytelnik mógł sobie wyrobić pogląd na temat zgodności analiz i interpretacji autora opracowania z surowymi danymi językowymi. Niemniej, oczywistą rzeczą jest (z której trzeba zdawać sobie sprawę), że taka prezentacja danych ściśle językowych nie była celem samym w sobie monografii Gralińskiego. Celem było, jak sam pisze we *Wstępie* (a dokładniej: w *Przedmowie*), omówienie szerokiego spektrum badań możliwych do przeprowadzenia przy użyciu narzędzia „Odkrywka" – od językoznawstwa po informatykę, folklorystykę i bibliotekoznawstwo. Zatem, dodaje dalej Graliński, w zależności od tego,

kim jest czytelnik, można zastosować różne strategie czytania, jednak nie zakłada się, że czytelnik musi mieć wysokie kompetencje informatyczne (z oczywistym wyjątkiem informatyków!), ponieważ wszystkie pojęcia informatyczne są wyjaśnione, a części informatyczne książki są całkiem samodzielne, odrębne. Autor zwraca także uwagę, że większość opisanych tu prac dotyczy projektu „Odkrywka", zatem czytelnik dowie się – rozdział po rozdziale – jak narodziła się „Odkrywka" i jak ją można wykorzystać do uzyskania wglądu w język polski i jego historię.

Ten cel postawiony na samym początku monografii został w pełni przez autora zrealizowany. Czytelnik dostaje do ręki opracowanie bardziej metodologiczno-teoretyczne, zanurzone w teoriach kilku dyscyplin, dziedzin i specjalności naukowych, niż opracowanie ściśle materiałowe, w którym zaprezentowano by wiele szczegółowych analiz językoznawczych (one oczywiście też się pojawiają, jako *case studies*). To, jakiego rodzaju analizy językoznawcze można przeprowadzić za pomocą opisanego w monografii narzędzia cyfrowego „Odkrywka", jak wysokiej jakości i precyzji mogą być to analizy struktur języka polskiego, możemy się na podstawie opisanych przez autora metodologii domyślać.

Na koniec należy dodać jedno spostrzeżenie natury kompozycyjnej. We *Wstępie* autor zapowiada bardzo dokładnie, co zostanie w monografii opisane, jakie zagadnienia zostaną rozpatrzone. W kontekście szczegółowych rozważań i analiz prowadzonych w poszczególnych rozdziałach (i częściach) książki nieco brakuje rozdziału końcowego, podsumowującego i wskazującego – potencjalne, w ujęciu autorskim – dalsze badania, ich perspektywy, czy wpływ prac(y) Filipa Gralińskiego na „zależne" od monografii dyscypliny, dziedziny, specjalności wiedzy. Od pracy omawiającej tak pionierskie narzędzie można by oczekiwać podsumowania będącego swoistym planem badawczym na przyszłość. Tego czytelnik może się – znów – jedynie domyślać po lekturze ostatniej, czwartej części: *APPLICATIONS*, wnioskując jednocześnie z całości monografii, jakie są możliwe drogi badawcze autora i potencjalny rozwój tych kilku dyscyplin, dziedzin, specjalności wiedzy pozostających w analogicznym obszarze badań.

Wskazane powyżej drobne mankamenty nie umniejszają, oczywiście, sygnalizowanego już wcześniej bardzo dużego znaczenia monografii Filipa Gralińskiego dla polskich badań z pogranicza lingwistyki i informatyki. Naukowa otwartość jego dzieła przejawia się także w podejmowaniu działań interdyscyplinarnych. Trzeba bowiem dodać, że przecież język naturalny z punktu widzenia matematyki jest traktowany jako struktura nieciągła, dyskretna (co *notabene* nie jest wielkim zaskoczeniem dla lingwistów). Jak pisze Graliński: „Język jest dyskretny, ale czas jest ciągły, a interakcja

ciągłości i dyskretności sprawia, że budowanie czasowych modeli językowych jest trudnym zadaniem, zarówno teoretycznie, jak i praktycznie" (s. 153). Wysoka świadomość metodologiczna, wiedza oraz kompetencje badawcze umożliwiły umiejscowienie analizowanego zagadnienia w kontinuum naukowo-poznawczym: od dyskretności podejścia matematycznego do nieciągłości podejścia językoznawczego. Należy zatem powtórzyć z całą mocą, że recenzowana monografia Filipa Gralińskiego *Against the Arrow of Time. Theory and Practice of Mining Massive Corpora of Polish Historical Texts for Linguistic and Historical Research*, opisująca stworzone przez niego narzędzie korpusowe (system „Odkrywka") i tkwiące w tym narzędziu dalsze możliwości badawcze istotne dla takich dyscyplin, jak informatyka, lingwistyka, folklorystyka, ocenić trzeba jako osiągnięcie klasy światowej.