dr hab. Magdalena Derwojedowa
Instytut Języka Polskiego
Wydział Polonistyki
Uniwersytet Warszawski
Krakowskie Przedmieście 26/28
00-729 Warszawa
derwojed_at_uw.edu.pl

Recenzja dorobku naukowego dra Filipa Gralińskiego w postępowaniu habilitacyjnym przeprowadzanym przed Senatem Uniwersytetu im. Adama Mickiewicza w Poznaniu

1. Uwagi wstępne

Doktor Filip Graliński od początku swojej kariery naukowej przejawia zainteresowania przetwarzaniem języka naturalnego (*natural language processing*, NLP). Zgłoszona jako główne osiągnięcie naukowe Habilitanta rozprawa *Against the Arrow of Time. Theory and Practice of Mining Massive Corpora of Polish Historical Text for Linguistics and Historical Research* wydana przez Wydawnictwo Naukowe UAM w 2019 r. jest kolejnym krokiem na drodze rozwoju tych zainteresowań. Zapoczątkowała je praca nad przekładem maszynowym, praca doktorska poświęcona była trudnemu zagadnieniu nieciągłości składniowej, przedstawiana dysertacja habilitacyjna dotyczy wydobywania danych językowych (*data mining*).

Monografię wskazaną jako główne osiągnięcie Habilitanta uzupełnia siedem artykułów, z których pięć to prace współautorskie (por. pkt 2), w części poświęcone tej samej tematyce, co rozprawa. Pozostałe z nich dotyczą zagadnień prezentujących szerokie zainteresowania dra Gralińskiego: zastosowanie uczenia maszynowego do usuwania błędów w danych i modelach wykorzystywanych w przetwarzaniu języka naturalnego, narzędzia do rywalizacji i współpracy w rozwiązywaniu zadań z NLP (Gonito.pl) oraz badania folklorystyczne, które dały zaczątek stanowiącemu główne osiągnięcie Habilitanta systemowi Odkrywka. Opublikowany w 2018 r. artykuł *Odkrywka, czyli leksykografia diachroniczna live* (współautorstwa Piotra Wierzchonia) stanowi podstawę rozdziału w dysertacji (o czym Autor informuje we wstępie do książki).

2. Główne osiągnięcie

Rozprawa *Against the Arrow of Time. Theory and Practice of Mining Massive Corpora of Polish Historical Text for Linguistics and Historical Research* (Wydawnictwo Naukowe UAM, 2019 r.) jest głównym osiągnięciem naukowym wskazanym przez Habilitanta.

Jest to opracowanie średniej objętości (15,75 ark. wyd.), przedstawiające założenia, metody i narzędzia badawcze dziedziny, którą można szeroko określić jako komputerowe językoznawstwo diachroniczne, węziej jako diachroniczną leksykografię komputerową, a najściślej — terminu tego używa sam Habilitant — jako lingwochronologizację, czyli „procedurę przyporządkowania informacji chronologizacyjnej jednostkom językowym"[1]. Rozważane zagadnienia Autor ilustruje przykładami z badań, w których uczestniczył, mamy więc do czynienia nie tylko z planem badawczym, ale też z praktyczną weryfikacją teorii. Jest to w moim przekonaniu znacząca zaleta monografii. Przykłady pozajęzykoznawczego wykorzystania wydobytych metodami automatycznymi danych językowych (ang. *data mining*) w naukach humanistycznych (reprezentowanych przez kulturoznawstwo i folklorystykę) uważam za dodatkowo podnoszące wartość monografii F. Gralińskiego.

Sama rozprawa może być przewodnikiem dla humanistów (dla osób zorientowanych nieco w metodach ilościowych interesujący i ogólnie zrozumiały będzie także bardziej techniczny rozdział 6., co do którego Autor czyni bezpieczne zastrzeżenie), jak i informatyków zainteresowanych przetwarzaniem tekstów — pierwszym przybliży narzędzie, jakim jest system Odkrywka oraz inne narzędzia używane do badania rozległych zasobów tekstowych, dla drugich będzie kompetentną dokumentacją systemu projektowanego do pozyskiwania i opracowywania danych oraz pracy na tak uzyskanym materiale.

Książka opatrzona została dość bogatym materiałem ilustracyjnym (blisko 100 ekscerptów), dość obszerną bibliografią (około 120 pozycji), wykazami tabel, wykresów i ekscerptów fotograficznych oraz — co w polskich monografiach humanistycznych nie jest akie częste — skorowidzem. Monografię wyróżnia dość ekscentryczny (a utrudniający czytanie) układ typograficzny oraz żywy, obrazowy język nowoczesnych dzieł naukowych.

Dzieło dra Gralińskiego składa się z jedenastu rozdziałów podzielonych na cztery części, poświęconych kolejno: 1) tekstom, czyli materiałowi badawczemu, oraz jego opracowaniu; 2) poszukiwaniom badawczym (zgrabnie ujętym w formule *(re)searching*) w wyszukanych w materiale danych; 3) modelowaniu temporalnemu języka (głównie w aspekcie zasobu leksykalnego w danym przedziale czasowym) i 4) przykładowym zastosowaniom stworzonego przez Habilitanta systemu Odkrywka.

Jak informuje *Foreword*, cztery rozdziały oraz części dwóch kolejnych zostały oparte na wcześniej publikowanych artykułach w czasopismach lub publikacjach zbiorowych. Pięć z nich to prace współautorskie, tworzone w gronie poznańskich informatyków zajmujących się przetwarzaniem języka naturalnego, lub we współpracy ze środowiskiem lingwochronologów z rodzimego uniwersytetu Habilitanta, przede wszystkim z Piotrem Wierzchoniem. Wypada zatem przyjąć, że nieco ponad połowa monografii została opracowana na nowo, jest to większość części I i II oraz po jednym rozdziale z części III i IV. Mamy zatem do czynienia z tomem łączącym opracowanie oryginalne z opra-

[1] P. Wierzchoń, F. Graliński, *Odkrywka, czyli leksykografia diachroniczna live*, [w]: M. Bańko, H. Karaś, (red). *Między teorią a praktyką. Metody współczesnej leksykografii*, t. 1, Warszawa: Wydawnictwa Uniwersytetu Warszawskiego; 2018, s.59–69.

fologicznej niż Morfologik, jeśli ten obarczony jest tak poważnym — z punktu widzenia analizy językowej — błędem jak lematyzowanie obcego skrótu TV jako TELEWIZOR, ani nie ma możliwości zgadywania lematów form nieznanych, szczególnie że narzędzia podobne istnieją (np. analizator Morfeusz Marcina Wolińskiego). Wybór lematyzatora jest oczywiście decyzją badacza, zabrakło mi jednak dyskusji, z jakich narzędzi może on wybierać i uzasadnienia ostatecznej decyzji.

Nie przekonuje mnie „normalizacja ortograficzna" wyszukiwania, choć jest to rozwiązane popularne w przetwarzaniu zasobów diachronicznych (np. podobne nieszczęśliwe rozwiązanie zastosowano w wyszukiwarce korpusu XVII i I poł. XVIII w. KorBa). W ten sposób gubi się wiele interesujących informacji, np. w przebiegach czasowych nie jest widoczne, kiedy zmianie uległa pisownia czy odmiana. Z danych cząstkowych znaczenie łatwiej jest wyliczyć przebiegi sumaryczne niż w „ujednoliconych" danych wyszukiwać formy odmienne od „znormalizowanych" (czyli w praktyce współczesnych). Jest to zapewne uboczna cecha systemu nastawionego na wyszukiwania czysto leksykalne, jednak językoznawstwo ma też inne cele, np. śledzenie zmian w systemie fleksyjnym czy składniowym. Absolutyzację informacji leksykalnej wyraziście ilustruje wykres końcówek narzędnika i miejscownika przymiotników (rys. 4.4), na którym informacja o ich rozkładzie została stracona (wbrew temu, co pisze Autor, wcale nie wariantywnych). Jak istotna może być informacja tego typu pokazuje Habilitant w rozdziale 7., kiedy opisuje metody automatycznej klasyfikacji temporalnej tekstów (7.4 i rys. 7.3).

Mimo potencjalnie nieograniczonego zasięgu czasowego, prezentowane wyniki z Odkrywki pochodzą w większości z tekstów po 1800 r., wyjątkiem jest rozdział 8., gdzie pojawiają się wykresy dotyczące lat 1600-2000. Z jednej strony pozwala to w jakimś stopniu porównać wyniki z Odkrywki z wynikami z Google nGram, z drugiej — dotyczą one doby nowopolskiej, dla której istnieją obszerne słowniki i opracowania, a także swobodnie dostępne, obszerne i dynamicznie przyrastające zasoby tekstów w bibliotekach cyfrowych, przynajmniej częściowo umożliwiające weryfikacji charakterystyki chronologicznej wyrazów.

Będące częścią systemu Odkrywka środowisko pracy *dossier* ułatwia zapewne pracę z wyszukanymi danymi, nie odbiega jednak od codziennej praktyki weryfikacji przykładów z konkordancji — potwierdzania wątpliwych i odrzucania błędnych. Niewątpliwie nowy standard na gruncie polskim stanowią łatwe do uzyskania wykresy chronologiczne, jest to standardowa funkcja nowoczesnych serwisów korpusowych (np. korpusów DWDS czy współczesnego polskiego korpusu monitorującego MonCo autorstwa Piotra Pęzika).

Część III, Modelling, składa się z trzech rozdziałów. To bez wątpienia najbardziej interesująca i najlepsza część rozprawy. Rozdział 6. zawiera objaśnienie formalnych metod użytych do modelowania języka z uwzględnieniem jego charakterystyki w zadanych przedziałach czasowych. Choć rozdział ten jest najbardziej techniczny i skierowany głównie do inżynierskiego odbiorcy monografii, jest napisany bardzo przejrzyście i korzyści z niego odniosą nawet czytelnicy nie czujący się pewnie w (nie aż tak licznych) podanych w nim wzorach. Kolejny rozdział na przykładach przedstawia, jak można wykorzystać modele temporalne języka do konkretnych, również bardzo praktycznych zastosowań. Za bardzo instruktywne i poznawczo ciekawe uważam części prezentujące korpus treningowy i prostsze i bardziej wyrafinowane metody wykorzystane do ustalania czasu powstania tekstów. Bardzo interesujący i poznawczo cenny jest także rozdział 8., poświęcony wykrywaniu synonimów.

cowaniem problemów wcześniej dyskutowanych i opublikowanych — co jest praktyką znaną w literaturze naukowej na świecie, a stosunkowo rzadszą na gruncie polskim. Znaczącą zaletą tego rozwiązania jest to to, że czytelnik zyskuje kompletny obraz omawianych zagadnień w jednolitej publikacji.

Podstawowym celem książki jest przedstawienie zasad wydobywania danych językowych z tekstów dostępnych pierwotnie (ang. *digital-born*) lub wtórnie (ang. *digitalized*) w postaci zdygitalizowanej. Jej podstawą lingwistyczną jest teoria lingwochronologizacji sformułowana przez Piotra Wierzchonia.

Na wstępie monografii F. Graliński formułuje dwanaście dyrektyw tworzenia zasobu do badań lingwochronologicznych. Część z nich ma charakter czysto technicznym, część — ideowy. Reguły D1 i D2 postulują utworzenie jednego repozytorium, które mogłoby być rozbudowywane przez każdego z użytkowników, oraz stworzenie metawyszukiwarki do wszystkich zasobów. Reguły D3-D8 omawiają uwzględnione (możliwe do relatywnie łatwego pozyskania) źródła tekstowe, reguły D9-D12 stanowią, że proces wyszukiwania, archiwizowania i przygotowywania fotodokumentacji powinien być zautomatyzowany i uniwersalny, tzn. powinno być możliwe tworzenie i gromadzenie zasobów wielojęzycznych. Wszystkie postulaty można sprowadzić do trzech: 1) elektronicznego repozytorium zasobów; 2) umieszczenie w nim jak największego zbioru tekstów danego języka w postaci pisanej i stała aktualizacja zasobu; 3) metawyszukiwarki. Proponowany przez Habilitanta korpus nie jest sam w sobie *novum*, jednak od rozwiązań na gruncie polskim różni go przede wszystkim podejście — w założeniu totalne, czyli objęcie oglądem tekstów bez cezury czasowej czy stylistycznej, a w praktyce realistyczne, a więc przeszukiwanie zasobów, które można zgromadzić lub do których można mieć dostęp, choć niemożliwe jest ich zgromadzenie (jak np. archiwa prasowe). Jednocześnie zakłada on nieustanną rozbudowę korpusu nie tylko o teksty najnowsze, ale też o przybywające w ostatnich latach coraz szybciej teksty dawne.

Takie narzędzie ma służyć wyszukiwaniu wystąpień zadanych ciągów (o dowolnym statusie językowym) i przypisywaniu im informacji chronologicznej, tzn. maksymalnie dawnego datowania potwierdzonego w korpusie i przebiegu czasowego wystąpień w tekstach. O ile samo pozyskiwanie, porządkowanie i udostępnianie danych zostało zautomatyzowane, o tyle ich interpretacja zakłada (słusznie) udział świadomego badacza, który ma do dyspozycji środowisko *dossier* ułatwiające przeglądanie i ewaluowanie materiału.

Części I i II poświęcone są technicznym zagadnieniom gromadzenia i opracowania tekstów oraz pracy z serwisem Odkrywka. Rozdziały te, choć nie przynoszą rewelacji lingwistycznych nawet na małą skalę, bardzo dobrze pokazują trudności, z jakimi musi się zmierzyć każdy twórca większego zasobu językowego — od niewyraźnych tekstów po spójne metadane. Uwagi Habilitanta o standaryzacji metadanych i prezentację metod ich ujednolicania uważam za wzorcowy przykład nowoczesnego podejścia do rzetelnej dokumentacji faktów językowych wart polecenia wszystkim prowadzącym badania materiałowe. Ciekawy poznawczo jest też opis problemów, jakie nastręcza rozpoznawanie pisma (*optical character recognition*, OCR), szczególnie w starszych publikacjach. W tej części Autor wyjaśnia też nieduży zasięg czasowy gromadzonego korpusu. Trzeba jednak zauważyć, że dopiero teksty dawniejsze stanowią obszar rzeczywiście niezbadany metodami ilościowymi, a wymagającym nie tylko zmian w rozpoznawaniu pisma, ale też odpowiednio przygotowanych lematyzatorów i tagerów. Sądzę natomiast, że niektórych niedogodności można było uniknąć, np. wybierając inny moduł analizy mor-

to dwie jednostki, jeśli natomiast notujemy pierwsze pojawienie się pewnego ciągu (unilateralnego), uzasadnienie tracą „warianty historyczne" pisowni — są one bowiem osobnymi jednostkami i ekwiwalentami diachronicznymi.

Za niedoskonałość opracowania należy też uznać nieumieszczenie niektórych pojęć w skorowidzu (np. *chronizm*, ang. *chronism*), a włączenie do niego nazwisk wszystkich cytowanych autorów, a nawet hasła *Niepokalanów*, raczej nieistotnego ze względu na treść rozważań. Do wad skorowidza należy też zaliczyć to, że prowadzi on do pozycji bibliograficznych, a nie do definicji pojęć (jak lingwochronologizacja ang. *linguochronologisation*) lub do wzmianek — tak jest w przypadku jednostki językowej (*linguistic unit*) czy neologizmu (*neologism*).

Braki te jednak nie wpływają na moją zdecydowanie pozytywną ocenę dysertacji Habilitanta. Prezentuje ona spójną i konsekwentną postawę badawczą dra Filipa Gralińskiego, wyczerpująco i przekonująco uzasadnia wykorzystane metody i narzędzia, przedstawia wreszcie znaczącą liczbę ciekawych przykładów pozyskanych przez Autora za pomocą zaproponowanej przez niego metody.

3. Inne publikacje i osiągnięcia naukowe

Artykuły *GEval: Tool for Debugging NLP Dataset and Models* (współautorstwa A. Wróblewskiej i T. Góreckiego) i *Gonito.net — Open Platform for Research Competition, Cooperation and Reproducibility* (współautorstwa R. Jaworskiego, Ł. Borchmanna i P. Wierzchonia) mają charakter techniczny i dotyczą narzędzi informatycznych z zakresu NLP, których (współ)twórcą jest Habilitant. Do tej grupy prac należy też zaliczyć *Mining the Web for Idiomatic Expressions Using Metalinguistic Markers*, w którym został przedstawiony pomysłowy sposób ekstrakcji idiomów z tekstów.

Artykuł *Z historii „parcia" na neologizmy* (współautorstwa P. Wierzchonia) można potraktować jako tekst programowy. Zawiera on wiele przykładów redatacji neologizmów z bazy Obserwatorium Językowego UW, dobrze udokumentowanych. Poza oczywistym pożytkiem ze sprostowania błędnych danych, jego główną wartością jest w moim przekonaniu krytyczne, zwięzłe, omówienie sześciu stosowanych w literaturze kryteriów neonimiczności, opatrzone przykładami.

Artykuły *Odkrywka, czyli leksykografia diachroniczna* live (współautorstwa Piotra Wierzchonia) i *System Odkrywka jako innowacyjne narzędzie informatyczne do badania polskiej leksyki potocznej. Przykłady zastosowania* (współautorstwa D. Dzienisiewicza i K. Świetlika) przedstawiają program Odkrywka i jego główne cechy użytkowe, uzupełnione o przykłady zastosowań do datowania na materiale *Słownika polskich leksemów potocznych* W. Lubasia. Szczególnie druga pozycja pokazuje wiele interesujących przypadków (re)datacji i jest przekonującym argumentem za wykorzystaniem proponowanych przez dra Gralińskiego metod do datacji jednostek leksykalnych.

W artykule *Folklorystyka 2.0* Habilitant prezentuje się jako badacz tzw. folkloru miejskiego i formułuje założenia korpusu, który ma posłużyć od śledzenia tzw. legend miejskich, oraz narzędzi do jego gromadzenia i przeszukiwania. Założenia te stały się (dosłowną) podstawą dyrektyw sformułowanych w dysertacji.

Artykuły dotyczące diachronii stosunkowo niewiele poszerzają problematykę ujętą w monografii, są jednak świetnym świadectwem rozwoju zainte-

Co prawda sama metoda jest znana od dawna (*latent semantics analysis*, LSA), a na gruncie polskim w została wykorzystana przez twórców Słowosieci 1.0 do konstruowania funkcji podobieństwa[2], ale użycie jej do badań diachronicznych stanowi w językoznawstwie polskim *novum*. Filip Graliński umiejętnie wprowadza czytelnika w interpretację danych liczbowych i instruktywnie prezentuje możliwości i pułapki zastosowanego aparatu. Za szczególnie wartościowe uważam przykłady pokazujące nieoczywiste podobieństwo, takie jak wyparcie patefonu przez gramofon, a tego ostatniego przez magnetofon czy przywołująca skojarzenie ze sławnym badaniem Pierre'a Guiraud „podobieństwo" między przedwojennym premierem Aleksandrem Prystorem a byłą premier Beatą Szydło.

W części IV Habilitant omawia kilka przykładowych badań, w których wykorzystał możliwości zbudowanego przez siebie systemu. Najciekawsze z punktu widzenia językoznawczego jest wykrywanie efemeryd leksykalnych bazujące na analizie przebiegów czasowych. Kolejne badanie ma charakter porównawczy i dotyczy tropów kulturowych, zestawia ono dane z Odkrywki z wynikami z Google nGram. Ostatni przykład to analiza obiegowych historii z dreszczykiem; wykorzystanie materiału zgromadzonego w korpusie ze starannie opracowanymi metadanymi umożliwia nie tylko prześledzenie przebiegu czasowego, ale też zasięgu geograficznego. Bogata ilustracja materiałowa tej części pokazuje dobitnie, jakie korzyści przynosi wykorzystanie sprawnych narzędzi do wydobywania danych (*data mining*) i jest bardzo mocny argumentem za proponowanym przez Habilitanta podejściem badawczym oraz stworzonym przez Niego narzędziem.

Ciekawa i pożyteczna dysertacja Filipa Gralińskiego ma też usterki, które jednak nie podważają jej ogólnie wysokiej wartości. Za zasadniczy i główny niedostatek rozprawy uważam swobodne i nieprecyzyjne posługiwanie się pojęciami dotyczącymi opisywanych zjawisk językowych. Brakuje przede wszystkim wyczerpującej definicji jednostki leksykalnej (*lexical unit*), podstawowa dla rozprawy *lingwochronologiznacja* odsyła do artykułu napisanego z P. Wierzchoniem. Sądzę, że byłoby z korzyścią, gdyby wszystkie ważniejsze pojęcia zostały w rozprawie formalnie zdefiniowane, np. efemeryda odwołuje się do neologizmu, ale na jakiej podstawie ustalić neonimiczność pewnego ciągu? Pojęcie jednostki leksykalnej wydaj mi się niezbędne, gdy ustalana jest ekwiwalencja diachroniczna, np. *komputer* to zarazem *mikrokomputer*, jak i *mózg elektronowy*, nie ma jednak, jak sądzę, ekwiwalencji między *mikrokomputerem* a *mózgiem elektronowym*. Czy w takim razie mamy do czynienia z homonimicznymi jednostkami KOMPUTER$_1$ i KOMPUTER$_2$? Podobne wątpliwości budzą *aeroplan* i *samolot*, a także używana w XIX w. w tym znaczeniu *aerodyna* (dziś o szerszym znaczeniu). Dla odmiany ciąg *polski papież* przed 1978 r. dowodzi tylko, że jednostka o tej postaci graficznej była składnikiem innego ciągu synonimicznego i nie mamy do czynienia z rewizją datowania, ale ewentualną homonimią. (Osobną sprawą jest, czy *polski papież* nie jest ciągiem charakteryzującym wyłącznie lub głównie dosłowny ekwiwalent przekładowy w przekładach z innych języków, nie funkcjonuje natomiast w tekstach pierwotnie polskich). Jeśli bowiem datowanie dotyczy jednostki bilateralnej, są

[2] Broda, B., M. Derwojedowa, M. Piasecki i S. Szpakowicz, *Corpus-based Semantic Relatedness for the Construction of PolishWordNet*, „Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)", European Language Resources Association (ELRA), 2008; Broda, B. i M. Piasecki, *Parallel, Massive Processing in SuperMatrix – a General Tool for Distributional Semantic Analysis of Corpora*, International Journal of Data Mining, Modelling and Management, 2011

resowań badawczych dra Gralińskiego, dowodzą jego obecności w polskim obiegu językoznawczym i folklorystycznym. Artykuły *stricte* informatyczne przekonują, że Habilitant nie tylko śledzi na bieżąco rozwój NLP, ale też uczestniczy w tworzeniu nowoczesnych narzędzi i doskonaleniu metod tej ważnej współcześnie dziedziny.

4. Konkluzja

Po zapoznaniu się z dorobkiem naukowym dra Filipa Gralińskiego stwierdzam, że spełnia on wymagania art. 16 ust. 2 ustawy z dnia z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki z późn. zm. (Dz. U. z 2017 r. poz. 1789) i wnoszę o dopuszczenie Habilitanta do dalszych etapów postępowania.

19 maja 2020 r.

Magdalena Derwojedowa

Warszawa, 19 maja 2020 roku