dr hab. Michał Szczyszek
Instytut Filologii Polskiej UAM w Poznaniu

Recenzja
dorobku naukowego, organizacyjnego i dydaktycznego
dra Filipa Gralińskiego
w związku z postępowaniem o nadanie stopnia doktora habilitowanego

Doktor Filip Graliński ukończył informatyczne studia magisterskie w 2001 r. na Wydziale Matematyki i Informatyki Uniwersytetu im. A. Mikiewicza w Poznaniu (temat pracy magisterskiej: Komputerowa analiza tekstu polskiego. Zastosowanie w systemie POLENG). Stopień doktora nauk matematycznych w zakresie informatyki uzyskał w 2007 r. także na Wydziale Matematyki i Informatyki Uniwersytetu im. A. Mikiewicza w Poznaniu na podstawie rozprawy *Formalizacja nieciągłości zdań przy zastosowaniu rozszerzonej* gramatyki bezkontekstowej napisanej pod kierunkiem prof. dra hab. Zygmunta Vetulaniego.

Działalność zawodowa dra Filipa Gralińskiego – jako nauczyciela akademickiego i pracownika naukowego zatrudnianego w jednostkach naukowych – związana jest przede wszystkim z Uniwersytetem im. Adama Mickiewicza w Poznaniu, na którym w okresie od października 2008 do września 2018 był zatrudniony na etacie adiunkta – na Wydziale Matematyki i Informatyki, a od października 2018 – na etacie starszego wykładowcy (też na Wydziale Matematyki i Informatyki).

1. Ocena aktywności naukowej

Dorobek naukowy dra Filipa Gralińskiego obejmuje 32 publikacje autorskie i współautorskie (tu: ze znacznym wkładem Habilitanta – w opracowanie zagadnienia naukowego, przygotowanie artykułu – opisanym procentowo przez Niego w załączniku autobibliograficznym do Jego wniosku habilitacyjnego). Wśród tych prac – wydanych po uzyskaniu stopnia doktora (najstarsza z nich datuje się na 2009, najnowsza – na 2019 r., a 2 artykuły przyjęte do druku czekają na publikację) – znajdują się: 1 autorska monografia (stanowiąca podstawę postępowania habilitacyjnego), 31 artykułów lub rozdziałów w

monografiach wieloautorskich – w tym 3 artykuły opublikowane w czasopismach znajdujących się w bazie WoS lub na liście ERIH. Artykuły autorstwa Habilitanta to teksty autorskie i współautorskie, pisane zarówno po polsku, po rosyjsku, jak i po angielsku (wydawane m. in. w takich czasopismach i wydawnictwach, jak np. „Poradnik Językowy", „Studia Sociologica", „Studia Rossica Gedanensia", Wydawnictwo Uniwersytetu Warszawskiego, Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu, czy Springer Berlin Heidelberg i inne). Sumaryczny impact faktor (wg JCR) Habilitanta wynosi: 0,25; liczba cytowań (wg WoS) – 8, a indeks Hirscha (wg WoS) – 2. Pod względem ilościowym jest to dorobek aż nadto wystarczający, a przy tym warto podkreślić, że wartościowy oraz spójny merytorycznie i metodologicznie i nieustannie rozwijany, co pokazują wydawnictwa i czasopisma, które opublikowały lub przyjęły do druku opracowania Habilitanta.

W kręgu zainteresowań Habilitanta od początku Jego drogi naukowej pozostają – jak Habilitant sam pisze w autoreferacie (s. 9 Autoreferatu) – trzy obszary badawcze: informatyka (z nastawieniem na lingwistykę komputerową, czy inaczej mówiąc: wykorzystywanie narzędzi cyfrowych w humanistyce oraz przetwarzanie języka naturalnego), językoznawstwo oraz folklorystyka. Analizując Jego dorobek naukowy – zwłaszcza ten, po uzyskaniu stopnia doktora – nie sposób temu zaprzeczyć (ponadto na informatyczne zainteresowanie lingwistyką wskazują tematy Jego pracy magisterskiej oraz dysertacji doktorskiej). Wszystkie te obszary badawcze wzajemnie się przenikają, są zależne od siebie. Habilitant, będąc z podstawowego swojego wyższego wykształcenia matematykiem, informatykiem realizuje badania na styku informatyki i etnologii/ folklorystyki (tu: w zakresie gromadzenia i badań tzw. legend miejskich) oraz na styku informatyki i lingwistyki (tu: opracowane przez Niego narzędzie i korpus językowy – system „Odkrywka" i badania w zakresie lingwistyki komputerowej, opracowywanie procedur np. z zakresu przetwarzania języka naturalnego (NLP)).

Można zatem powiedzieć, że wykorzystuje on swoją wiedzę i swoje kompetencje informatyczne do budowania narzędzi służących wyszukiwaniu danych etnologicznych/ folklorystycznych i lingwistycznych oraz budowaniu korpusów – zwłaszcza językowych. Jak sam zresztą pisze w autoreferacie – poszukiwania etnologiczne/ folklorystyczne mają u swoich podstaw dane, zjawiska, procesy językowe (por. s. 9 Autoreferatu). Zatem Habilitant doszedł do słusznego wniosku, że aby wykorzystać narzędzia cyfrowe w badaniach etnologicznych/ folklorystycznych, należy rozwijać narzędzia cyfrowe dla humanistyki,

skupiając się na narzędziach wypracowywanych w ramach lingwistyki komputerowej. Ten bardzo spójny metodologicznie, naukowo, teoriopoznawczo sposób myślenia badawczego i postępowania badawczego należy ocenić jako bardzo dobre podejście do rozwiązywania problemów naukowych. Habilitant mając kompetencje informatyczne, potrafi z bardzo dobrym skutkiem zastosować je w praktyce badawczej łączącej te trzy wymienione wyżej obszary badawcze. Stworzył bowiem bardzo operatywne narzędzie lingwistyczno-informatyczne: system „Odkrywka", który umożliwia tworzenie obszernego, nieograniczonego chronologicznie korpusu tekstów polskich. Należy zauważyć, że na drodze do stworzenia systemu „Odkrywka" Habilitant zdobywał doświadczenie teoretyczne i praktyczne, (współ)tworząc systemy do przetwarzania języka naturalnego (np. PSI-Toolkit: A Natural Language Processing Pipeline , 2013 i inne wymienione w Jego Autoreferacie).

Wysokie kompetencje naukowe Habilitanta oraz możliwości wypracowanej przez niego metody badawczej i narzędzia wyszukiwawczego (wraz z korpusem) widać wyraźnie z zbiorze Jego artykułów opublikowanych po uzyskaniu stopnia doktora, a dołączonych do wniosku habilitacyjnego. Teksty te – pisane zarówno po polsku, jak i o angielsku – dowodzą szerokich horyzontów badawczych. Przykładowo – w Jego publikacjach pojawiają się teksty dotyczące gromadzenia danych etnologicznych/ folklorystycznych – tak w wypadku zbioru legend miejskich: książka autorstwa Habilitanta: *Znikająca nerka. Mały leksykon współczesnych legend miejskich* (w innych tekstach z tego kręgu Habilitant pokazuje drogi badawcze i procedury gromadzenia danych etnologicznych/ folklorystycznych i ich interpretacji – por. np. Folklorystyka 2.0 autorstwa Habilitanta). Opracowania z zakresu lingwochronologii (por. tekst Odkrywka, czyli leksykografia diachroniczna live napisany przez Habilitanta we współautorstwie z P. Wierzchoniem), w tym nurcie – bardzo konstruktywna krytyka naukowa obserwatorium językowego UW (por. artykuł Z kart historii *„parcia na" neologizmy* napisany przez Habilitanta we współautorstwie z P. Wierzchoniem) oraz z zakresu lingwistyki kryminalistycznej (sądowej) (por. tekst *Przydatność tradycyjnych opisów leksykograficznych w dowodach z opinii biegłego językoznawcy dotyczących znieważenia. Na przykładzie rzeczownika „pedał" i w kontekście możliwości programu do* automatycznego wyszukiwania danych FBL Risercz napisany przez Habilitanta we współautorstwie z J. Liberkiem) pokazują i unaoczniają potencjał wykorzystywania narzędzi cyfrowych w badaniach językoznawczych w ujęciu diachronicznym i synchronicznym. Oczywiście, w ocenianym dorobku Habilitanta poczesne miejsce zajmują zagadnienia cyfrowe, informatyczne odnoszące się do zagadnień związanych przetwarzaniem języka

naturalnego (por. tekst Geval: Tools for debugging NLP Datasets and Models napisany przez Habilitanta we współautorstwie z Anną Wróblewską, Tomaszem Stanisławkiem), jednakże – zawsze przy zachowaniu wysokich wartości merytorycznych w zakresie informatyki i zawsze nastawionych na próbę rozwiązania problemów badawczych lingwistyki (cyfrowej).

Należy zauważyć z całym przekonaniem, że omawiane tu opracowania naukowe – powstające często we współpracy z lingwistami, jak i samodzielnie – są bardzo szczegółowe, dokładne, metodologicznie (informatycznie i lingwistycznie) spójne i konsekwentne. Wyniki badań prezentowanych w tych artykułach (np. tekst *Z kart historii „parcia na" neologizmy*) nie pozostawiają cienia wątpliwości interpretacyjnych, poznawczych. Są to teksty, w których bardzo dobrze wykorzystano teorie i metodologie naukowe (informatyczne i lingwistyczne), a rezultaty badawcze wydają się niepodważalne (np. por. tekst System Odkrywka jako innowa*cyjne narzędzie informatyczne do badania polskiej leksyki potocznej. Przykłady* zastosowania napisany przez Habilitanta we współautorstwie z Danielem Dzienisiewiczem, Karolem Świetlikiem). Należy stwierdzić z całą mocą, że każdy z omawianych tu tekstów wnosi nową wiedzę, nowe ustalenia do nauki, w różnych jej obszarach – oczywiście głównie w tych trzech obszarach działalności naukowej Habilitanta. Przykładowo – teksty System Odkrywka jako innowa*cyjne narzędzie informatyczne do badania polskie*j leksyki potocznej. *Przykłady zastosowania* czy Odkrywka, czyli leksykografia diachroniczna live są bardzo istotne dla procedury badań lingwochronologizacyjnych w zakresie języka polskiego (czyli ustalania daty pierwszego pojawienia się w tekstach danej jednostki leksykalnej (i szerzej: językowej), która to data zapewne stoi w pewnej bliskości pojawienia się danej jednostki leksykalnej w języku jako takim; ten aspekt działalności naukowej Habilitanta to oczywista kontynuacja badań i ustaleń zarówno Piotra Wierzchonia, jak i Jana Wawrzyńczyka). Warto jednakże dodać, że zawsze może się jednak okazać, że np. języku mówionym (idąc tropem Onga) niektóre wyrażenia językowe pojawiły się wcześniej niż możliwe do ustalenia ich wystąpienia w tekście pisanym; zdaje się, że tego typu refleksji (dotyczącej zróżnicowania języka mówionego i pisanego) nieco brakuje w świadomości badawczej Habilitanta. Nie należy oczywiście z tego robić szczególnego zrzutu, gdyż Habilitant deklaruje wszakże badania tekstów zachowanych w jakiejkolwiek postaci zapisanej za pomocą alfabetu (polskiego/ łacińskiego (czy potencjalnie innych)) i tu dopuszcza możliwość przesuwaniu datowania leksykalnego wraz z pojawieniem się nowo odkrytych tekstów dawnych, a zwłaszcza – z ich zdygitalizowaniem, a zwłaszcza – pełnotekstowym. Habilitant omawia także bardzo precyzyjnie problemy – czysto informatyczne – związane z procedurą

digitalizacji tekstów polskich, możliwością ich przeszukiwania za pomocą narzędzi cyfrowych (np. Mining the Web for idiomatic Expressions Using Metalinguistic Markers autorstwa Habilitanta). W ten sposób daje naukowe świadectwo znajomości problemów związanych z gromadzeniem danych językowych w sposób cyfrowy, proponując jednocześnie własną koncepcję rozwiązania tych problemów.

Jak widać – wszystkie obszary badawcze, po których poruszał się Habilitant – wzajemnie się w Jego pracach przenikają. Analizowane tu opracowania Jego autorstwa (lub z Nim jako współautorem) dają podstawy do stwierdzenia, że dorobek ten jest wysokiej próby. Jest to zbiór artykułów i opracowań pokazujący stale i konsekwentnie rozwijane zainteresowania badawcze Habilitanta, które zawsze oscylują wokół jasno sprecyzowanej osi, którą stanowi: informatyka, w tym informatyka lingwistyczna / lingwistyka komputerowa czy przetwarzanie języka naturalnego, czy jeszcze inaczej mówiąc – budowanie narzędzi i zasobów cyfrowych dla badań humanistycznych, zwłaszcza lingwistycznych (ale także i etnologicznych/ folklorystycznych). Tak szeroki wachlarz zainteresowań i badań nie spowodował swoistego rozproszenia teoretyczno-metodologicznego prac Habilitanta. Wręcz przeciwnie – wszystkie analizowane tu opracowania charakteryzują się wysokim stopniem konsekwencji badawczej, metodologicznej, a także akrybią (w tym zwłaszcza filologiczną), porządkiem argumentacyjnym i kompozycyjnym, wewnętrzną spójnością. Te uwagi można odnieść w całości do omawianego tu zbioru tekstów. Dowodzi on podjęcia się przez Habilitanta niełatwej drogi naukowej i realizowania jej w sposób bardzo konsekwentny, precyzyjny, niepodważalny i wewnętrznie spójny. Tylko w ten sposób – trzymając się ściśle teorii i metodologii naukowych można rozpatrywać złożone interdyscyplinarne problemy naukowe, a dorobek Habilitanta jest tego dobrym potwierdzeniem.

2. Ocena osiągnięcia naukowego, o którym mowa w art. 16 ust. 2 Ustawy

Zwieńczeniem obranej przez Habilitanta drogi badawczej jest zgłoszone do oceny osiągnięcie naukowe Against the Arrow of Time. Theory and Practice of Mining Massive Corpora of Polish Historical Texts for Linguistic and Historical Research obejmuje autorską książkę dra Filipa Gralińskiego Against the Arrow of Time. Theory and Practice of Mining Massive Corpora of Polish Historical Texts for Linguistic and Historical Research (Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu, Poznań 2019, ss. 315). Jest to kontynuacja i swoiste – aktualne, na tym etapie Jego drogi naukowej –

podsumowanie wcześniejszych rozważań i prac badawczych, zwłaszcza tych, które lokują się w dwóch Jego obszarach badawczych: 1) informatyka, zwłaszcza: zwłaszcza lingwistyka komputerowa i przetwarzanie języka naturalnego, 2) lingwistyka, zwłaszcza korpusologia i historia języka polskiego (doby nowopolskiej) – w tym: lingwochronologia. W ocenianej monografii natomiast, stanowiącej postawę postępowania habilitacyjnego dra Filipa Gralińskiego, na pierwszy plan wysuwają się zagadnienia związane z językoznawstwem (diachronicznym), lingwistyką komputerową, korpusologią i last but not least: lingwochronologią.

Praca ta jest bardzo cennym opracowaniem pokazującym problematykę, metodologię i procedury związane z budowaniem cyfrowych narzędzi dla humanistyki, 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. Habilitant swoje rozważania zawarte w monografii oparł na doświadczeniu i obserwacjach związanych ze stworzonym przez Niego narzędziu – systemie „Odkrywka". Narzędzie to gromadzi, przetwarza na podstawie stosownych algorytmów obszerne dane językowe z języka polskiego. Obszerność tych danych językowych zobrazować można informacjami o zasięgu chronologicznym: w korpusie są dane z całego okresu doby nowopolskiej i ze współczesności językowej, tj. od początków XIX wieku do dziś (por. s. 42 monografii); pojawiają się także (rzadziej, jeszcze nie w pełni systematycznie zbierane) dane z o okresu doby średniopolskiej (w niektórych wypadkach w „Odkrywce" pojawiają się dane nawet z 1600 roku i wcześniejsze). Zatem dysponujemy danymi językowymi obejmującymi kilkaset lat rozwoju polszczyzny (co najmniej od około 1800 roku, a niekiedy nawet od około 1600 roku (por. np. s. 75 monografii)) Jak pisze sam Habilitant, zbiór 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 kwantytatywnie wyrażoną wielkość korpusu „Odkrywka", to – zacytujmy monografię: „The total amount of plain text processed for Odkrywka – whether it be OCR-ed, manually transcribed or digital-born – is 96,368,642,437 characters (15,137,368,095 words – a word is defined as a continuous sequence of Unicodel letters and digits (...))" (s. 63). Ponad 15 miliardów wyrazów tekstowych (realizacji leksemowych) czyni z tego narzędzia – systemu „Odkrywka" – największy zbiór wyrazów dla języka polskiego (dla porównania – NKJP w wersji pełnej zawiera około 1 miliarda 800 milionów wyrazów) i

jeden z największych ze znanych i dostępnych korpusów światowych (może równać się ogromnymi korpusami języka angielskiego, np.: „NOW Corpus" („News on the Web", który – jak czytamy na stronie korpusu (https://www.english-corpora.org/now/): „contains 10.0 billion words of data from web-based newspapers and magazines from 2010 to the present time (the most recent day is 2020-05-20)") czy korpusem „iWeb corpus" (https://www.english-corpora.org/iweb/, który „contains 14 billion words (about 25 times the size of COCA) in 22 million web pages").

Należy oczywiście zauważyć od razu rzucającą się różnicę jakościową zachodzącą między korpusem stworzonym przez dra Gralińskiego, a przytoczonymi przykładowo korpusami języka angielskiego. Wynik porównania wypada na korzyść korpusu „Odkrywka". Na czym polega – w mojej, recenzenckiej ocenie – wyższość „Odkrywki" nad pozostałymi tu przytoczonymi korpusami? Korpus Habilitanta obejmuje swoim zasięgiem chronologicznym kilkaset lat rozwoju polszczyzny (o czym pisałem powyżej), a korpusy angielskie – około jednej dekady(!). Korpusy angielskie bazują na danych tylko internetowych, czyli „digital-born", natomiast korpus „Odkrywka" gromadzi dane „OCR-ed, manually transcribed or digital-born". Z danymi językowymi pochodzącymi z procesu digitalizowania (skanowania, poddawania OCR-owi) druków papierowych wiążą się kolejne niemałe problemy techniczne, metodologiczne, proceduralne, które Habilitant szczegółowo rozważał (o czym nieco szerzej – poniżej) i – rozwiązał!.

Na odrębną uwagę zasługuje także i to, że Habilitant w stworzonym przez siebie narzędziu i korpusie zgromadził 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, zwłaszcza wykresy na s. 22-23). Dla porównania – w przywołanych korpusach angielskich dane językowe są automatycznie ekscerpowane z internetowych stron anglojęzycznych lub z internetowych wydań gazet i czasopism anglojęzycznych.

Zatem – już teraz można z całą odpowiedzialnością powiedzieć, że opracowanie dra Filipa Gralińskiego, jak i jego narzędzie badawcze – „Odkrywka" – można (i należy) ocenić jako osiągnięcie naukowe klasy światowej w kategorii prac lingwistycznych, czy dokładniej – z zakresu lingwistyki komputerowej, korpusologii, przetwarzania języka naturalnego i lingwistyki. Oceniana praca oraz narzędzie i korpus w niej opisywane dają ogromne możliwości badawcze w zakresie np. wymienionej już wyżej lingwochronologii, a także w

zakresie 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 oficjalną), czy – last but not least – wyszukiwanie motywów, toposów, wątków literackich czy prasowych i etnologicznych (np. związanych z legendami (miejskimi) czy podaniami).

W odniesieniu do zgłoszonej do oceny monografii habilitacyjnej można więc z całą pewnością powiedzieć, że czytelnik otrzymuje wyważoną, logicznie spójną, precyzyjną i przejrzystą pod względem kompozycyjnym monografię 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* Habilitant 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 pierwszej: PART I. TEXTUAL MASSprzedyskutowane 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. Doktor Graliński, co trzeba z całą mocą powiedzieć, doskonale jest zorientowany we wszystkich aspektach cyfrowego gromadzenia danych językowych i przetwarzania języka naturalnego (NLP). Bardzo precyzyjnie wskazał źródła tekstów – pokazał, że wyzyskał materiał dostępny w

postaci zdygitalizowanej we wszystkich (bądź przynajmniej w większości) polskich bibliotekach cyfrowych, z uwzględnieniem tych najobszerniejszych i najbardziej rozpowszechnionych wśród użytkowników (jak Wielkopolska B.C.), 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 pod względem swoiście rozumianej ekstensji. Oznacza to, że w swoim korpusie zgromadził teksty należące do różnych odmian polszczyzny, reprezentujące różne gatunki wypowiedzi oraz różne stylistyki – do literatury pięknej po druki ulotne. Co więcej – ekscerpcja tekstów źródłowych z bibliotek cyfrowych dokonywana za pomocą opracowanego przez Niego narzędzia zasilającego korpus, to swoista never ending story, tj. wraz z dygitalizowaniem kolejnych druków, zasób korpusu „Odkrywka" będzie się zwiększał. Habilitant w monografii opisał także opracowaną przez siebie procedurę normalizacji metadanych np. 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. W ten sposób, dzięki tej procedurze dane zgormadzone w korpusie „Odkrywka" (ich datowanie) nie są narażone na swoiste „podróże w czasie", a zwłaszcza cofanie się do przeszłości. Rezultaty zastosowania procedur ustalania metadanych temporalnych zostały przekonująco opisane w monografii na kilku wybranych przykładach. Ostatnim punktem części pierwszej monografii jest opis działań zmierzających do uczynienia zdygitalizowanych druków czytelnymi dla narzędzi cyfrowych, czyli opis opracowanej samodzielnie i zastosowanej przez Habilitanta informatycznej metody związanej z NLP: (począwszy od „oczyszczenia" wyników pracy programu OCR, a skończywszy na przygotowaniu wyszukiwarki pełnotekstowej). Wszystkie te działania, opracowane procedury w zakresie NLP należy uznać za pionierskie na gruncie cyfrowego przetwarzania tekstów polskich (tj. w zakresie NLP w odniesieniu do języka polskiego); 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 Habilitanta).

W części drugiej monografii PART II. (RE)SEARCHING Habilitant przedstawił stosowaną przez siebie 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 . Autor pokazał bowiem w całej części drugiej swojej monografii zarówno procedurę pozyskiwania (wyszukiwania) jednostek językowych już z samego narzędzia – korpusu „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 popełnienia pomyłek w opisanych tu zakresach. W tej części monografii Autor przeprowadza swoisty instruktaż opisujący wszystkie wskazane tu aspekt pracy badawczej z „Odkrywką", unaoczniając i przestrzegając przed owymi pomyłkami. Przy okazji Habilitant prezentuje tu niemałą wiedzę i duże kompetencje w zakresie opracowywania i interpretowania danych statystycznych w odniesieniu do języka (polskiego).

Cześć trzecia PART III. MODELLINGwydaje się niezwykle ważna z punktu widzenia połączenia kompetencji informatycznych z lingwistycznymi dotyczącymi badań językoznawczych (zwłaszcza – historycznojęzykowych). Konieczność dostosowania modelu informatycznego (NLP) do danych historycznych języka polskiego wymusiła przyjęcie przez Habilitanta swoistego wyzwania związanego modelem ewaluacji danych historycznojęzykowych, przyjęcie określonego modelu samego języka w ujęciu diachronicznym czy wpracowania uczenia maszynowego dla tego typu danych. Zadanie niełatwe, wymagające niemałych kompetencji informatycznych i takowej wyobraźni badawczej, zwłaszcza że dr Graliński nie jest filologiem (diachronikiem), a właśnie informatykiem. Niekoniecznie wiec 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 punktu widzenia informatycznego (NLP) najlepiej wpisywały się w możliwości cyfrowe, technologiczne związane z lingwistyką komputerową właśnie. Wypracował zatem specjalne rozwiązanie RetroGapo o którym pisze „A machine learning challenge for temporal language models with log-loss hashed used as the evaluation metric was prepared and made available on the Gonito.net platform; see https://gonito. net/challenge/retro-gap. The data are easily accessible as a git repository at git://gonito.net/retro-gap.git. The challenge data sets were prepared using a general framework for generating diachronic challenges for the Polish language (cf. the RetroC(2) challenge for temporal classifiers later in this chapter as another example of such data sets, see Section 7.2). The challenge data were sampled from Odkrywka for the years 1814–2013. RetroGap corpora are divided into a training set, two development sets (dev-0 and dev-1) and a test set, as is usual for Gonito.net challenges; see Table 6.2" (s. 164) oraz korpus: RetroC corpus, o którym pisał: „RetroC is a Polish-language diachronic

corpus based on resources available in Odkrywka, spanning two centuries (1814–2013) and intended for training and testing automatic dating systems" (s. 177). I dalej doprecyzowywał: „There have been two releases of the corpus so far: the first (RetroC1) in 2015 and the second (RetroC2) in 2017. RetroC2 is not only larger (being a superset of RetroC1), but also contains extra features in the training set. The corpus was designed with the following goals in mind:
- to be a collection of Polish texts;
- to be large enough to enable the use of statistical methods;
- to be time-extensive – not just modern Web-based texts, but also old printed materials;
- to cover short fragments, not whole books (the dating task for whole books is much easier)". (s. 177).

To pozwoliło przyjąć „Word2vec model", o którym Habilitant szerzej pisze na s. 188 jako o zadaniu matematyczno-informatycznym. Nota bene – cała druga cześć monografii składa się z bardzo szczegółowych rozważań matematyczno-informatycznych w odniesieniu do istoty danych językowych, istoty języka.

Przy takim podejściu badawczym (zawierającym istotne elementy pragmatyki naukowej) Habilitant 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). Dowodzi to niemałych kompetencji poznawczych, naukowych, metodologicznych Habilitanta.

W części ostatniej, czwartej PART IV. APPLICATIONS Habilitant zaprezentował możliwości praktyczne związane z wykorzystaniem stworzonego przez siebie narzędzia (wyzyskiwanego z całym zapleczem teoretycznym, o którym pisze w części drugiej i trzeciej). Ta część to swoisty zbiór case studies związanych z jednej strony z poszukiwaniem efemerycznych struktur językowych (por. rozdział 9: Lexical ephemera) – a więc struktur znanych historykom języka też jako hapaks legomenon, a drugiej strony – folklorem (tu: miejskim), czy – jak pisze Habilitant – z folklorystyką 2.0 (odwołując się przy okazji do pojęcia kulturomics (por. rozdział 10: Traps of culturomics)). Nie wchodząc tutaj w dyskusję z dotychczasowymi próbami odnajdywana hapaks legomena (np. w słownikach notujących pewien procent „produkcji" językowej danego czasu), można pokusić się o przyjęcie założenia, że metody badawcze Habilitanta oraz możliwości tkwiące w Jego korpusie

pozwolą na faktycznie odnalezienie efemerycznych struktur językowych danego okresu rozwojowego polszczyzny. Wydaje się to osiągnięcie samo w sobie bardzo wartościowe, a w monografii pokazane na kilku przykładach – swoistych egzemplach. Podobnie rzecz się ma z możliwościami śledzenia wątków, „toposów" – tu: folklorystycznych. Habilitant 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. Cześć ta ma walor praktyczny: ukazania aplikacyjnych możliwości wynikających z prac Habilitanta. On sam doskonale zdaje sobie sprawę, czego wyraz dał właśnie w omawianych tu rozdziałach: jego osiągnięcia naukowe oraz system „Odkrywka" mogą przyczynić się do rozwoju takich dziedzin wiedzy, jak lingwistyka komputerowa, językoznawstwo polonistyczne (w ujęciu diachronicznym, synchronicznym, normatywnym), lingwochronologia, etnologia z folklorystyką (w tych aspektach badań etnologiczno-folklorystycznych, które bazują na materiale językowym).

Walorem pracy świadczącym o niebywałej kulturze naukowej Habilitanta jest ilościowo obszerna bibliografia. Bibliografia faktycznie wyzyskana na potrzeby monografii i dowodząca dużego oczytania Badacza w literaturze przedmiotu – w zakresie takich dyscyplin, dziedzin i specjalności, jak: informatyka, statystyka (matematyka), językoznawstwo (w różnych aspektach – od diachronii, przez synchronię, po zagadnienia kulturalnojęzykowe, normatywne), lingwistyka komputerowa, NLP (przetwarzanie języka naturalnego, a także i pośrednio: z zakresu sztucznej inteligencji), korpusologia, etnografia i folklorystyka, bibliotekoznawstwo (w tym – cyfrowe). Bibliografia ta pokazuje oczytanie Habilitanta w pracach – zarówno polskich, jak i europejskich czy światowych. Takiego oczytania i umiejętności zastosowania w swoich badaniach zastanej wiedzy, teorii i metodologii należy oczekiwać od habilitantów. W tym momencie należy ponownie podkreślić też świadomość metodologiczną dra Filipa Gralińskiego, który potrafił z tego bogactwa opracowań z różnych dziedzin z rozwagą wybrać odpowiednie metody, odpowiednio oraz twórczo zastosować (i dostosować) je do swoich potrzeb naukowych. Zapewniło to możliwość konfrontacji wyników przeprowadzonych badań z wcześniejszymi ustaleniami badawczymi w zakresie lingwistyki, językoznawstwa polonistycznego, przetwarzania języka naturalnego, lingwistyki komputerowej i korpusologii i innych dyscyplin wymienionych powyżej.

Mankamentem monografii może być to, że Habilitant nie prezentuje zbyt dużej liczby analiz ściśle językoznawczych. Przykładowo: 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 tak oczywiście, że tych analiz nie ma, że Habilitant nie pokazuje możliwości badań nad polszczyzną, zwłaszcza w ujęciu diachronicznym. Takie opisy, wykresy i ich interpretacje są, np. chronologia wystąpień wyrazu telewizja (por. s. 113); ponadto – cała cześć czwarta APPLICATIONS jest temu poświęcona. Można by w tym miejscu napisać, że od prac językoznawczych oczekuje się też przedstawienia – choćby w postaci aneksu, słownika, próbki danych językowych – zebranego materiału badawczego, aby czytelnik mógł sobie wyrobić pogląd na temat zgodności analiz i interpretacji autorów 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 pracy dra Gralińskiego. Celem Jego monografii było, jak sam pisze we *Wstępie* (a dokładniej: w Przedmowie): „This book may seem like a labyrinth, as a number of topics are covered here, from linguistics to computer science, folkloristics, and library science. Depending on who the reader is, various reading strategies can be applied. Nevertheless, no computer science competences are assumed (with the possible exception of Chapter 6) and you are encouraged to read the book in its entirety, as all notions are explained and the book is self-contained as far as computer science is concerned. Most of the work described here is related to the Odkrywka project, a linguistic experimental search engine operating on Polish historical texts. Chapter by chapter, you will learn how Odkrywka was born and how it (and its future successors) could be used to gain insight into Polish language and history". Ten cel postawiony na samym początku monografii (wyrażony zwłaszcza we fragmencie zaczynającym się słowami „Most of the work described here is related to the Odkrywka project....") 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 domyślać (ale – nie „wróżąc z fusów" metodologicznych, lecz precyzyjnie przewidując na podstawie opisanej przez Habilitanta metodologii i procedur). Zatem: możemy się domyślać i – spodziewać.

Na koniec należy dodać jedno spostrzeżenie natury kompozycyjnej. We *Wstępie* Habilitant zapowiada bardzo dokładnie, co zostanie w monografii opisane, jakie zagadnienie zostanie rozpatrzone. Nieco w tym kontekście oraz w kontekście mocno nasyconych treściami i rozważaniami Autora poszczególnych rozdziałów (i części) książki zabrakło rozdziału końcowego, podsumowującego te rozważania i wskazującego – potencjalne, w ujęciu autorskim – dalsze badania, ich perspektywy, czy wpływ prac(y) Habilitanta na „zależne" od monografii dyscypliny, dziedziny, specjalności wiedzy. Od tak pionierskiej pracy omawiającej tak pionierskie narzędzie (zrównane przeze mnie w niniejszej recenzji z pracami badawczymi na poziomie światowym) można by oczekiwać takiego podsumowania będącego swoistym planem badawczym na przyszłość. Tego czytelnik może się jedynie domyślać po lekturze ostatniej, czwartej części: APPLICATIONS wnioskując jednocześnie z całości monografii, jakie są możliwe drogi badawcze Autora i potencjalny rozwój tych kilku dyscypliny, dziedziny, specjalności wiedzy „zależnych" od monografii. Może o to Autorowi chodziło?

Wskazane powyżej drobne mankamenty nie umniejszają, oczywiście, sygnalizowanego już wcześniej bardzo dużego znaczenia pracy dra Filipa Gralińskiego. Habilitant jawi się tu jako świadomy, dojrzały badacz, który potrafi podjąć się złożonych badań (nad zjawiskami językowymi) z wykorzystaniem (meta)wiedzy i (meta)kompetencji informatycznych. Naukowa otwartość Go cechująca, przejawia się także w podejmowaniu działań interdyscyplinarnych – właśnie nad językiem naturalnym, a trzeba dodać, że z punktu widzenia matematyki język jest traktowany, opisywany jako struktura nieciągła, dyskretna (co nota bene nie jest to wielkim zaskoczeniem dla lingwistów); jak sam Habilitant pisze: „Language is discrete, but time is continuous, and the interaction of continuity and discreteness makes building temporal language models a challenging task, both theoretically and practically". (s. 153)). Wysoka świadomość metodologiczna, otwartość naukowa i wiedza wraz z kompetencjami badawczymi umożliwiły Habilitantowi umiejscowić analizowane zagadnienia, badaną problematykę w kontinuum naukowo-poznawczym; ma On umiejętności ekstrapolowania swoich ustaleń na szersze tło materiałowe i teoretyczne co stanowi niebagatelną zaletę Habilitanta. Należy zatem powtórzyć z całą mocą i odpowiedzialnością, że oceniane osiągnięcie naukowe dra Filipa Gralińskiego, jego praca badawcza, stworzone narzędzie i korpus (system „Odkrywka") ocenić trzeba jako osiągnięcie klasy światowej.

3. Pozostałe obszary aktywności naukowej, organizacyjnej i dydaktycznej

Doktor Filip Graliński bierze aktywny udział w życiu naukowym w Polsce i za granicą. Uczestniczył w 17 konferencjach krajowych (ogólnopolskich) i międzynarodowych (w tym i za granicami Polski), podczas których wygłosił interesujące referaty – samodzielne i współautorskie. Ponadto należy zauważyć dużą aktywność grantową dra Gralińskiego: na przestrzeni lat od 2011 do 2018 uczestniczył jako wykonawca w 7 różnych projektach badawczych finansowanych ze źródeł zewnętrznych (np. NPRH, MNiSW) realizowanych na UAM lub w kooperacji UAM – Samsung Electronics Polska. Jest to aktywność naukowa, której należałoby się spodziewać po habilitantach. Kooperacja w ramach przedsięwzięć grantowych Habilitanta z zewnętrzną firmą – Samsung Electronics Polska – jest niezwykłym świadectwem umiejętności Habilitanta współpracy z otoczeniem (tu: gospodarczym, biznesowym), co należy ocenić bardzo dobrze.

Doktor Graliński ma ponadto niemałe osiągnięcia na niwie dydaktycznej. Od 2001 roku (czyli od momentu rozpoczęcia Studiów Doktoranckich!) prowadził na rodzimym Wydziale Matematyki i Informatyki UAM zajęcia – wykłady i laboratoria – z przedmiotów informatycznych zorientowanych lingwistycznie, wpisujących się w Jego zainteresowania naukowe; zajęcia te, to: „Tłumaczenie maszynowe", „Inteligentne systemy informacyjne", „Automaty i języki formalne", Uczenie maszynowe" oraz last but not least! „Sztuczna inteligencja" (prowadzone przez Niego zajęcia dydaktyczne korelują z Jego zainteresowaniami naukowymi poświadczanymi np. tematem pracy magisterskiej i zagadnieniem opisanym w dysertacji doktorskiej oraz – co oczywiste i wyżej szczegółowo omówione – z aktywnością publikacyjną i konferencyjną). Ponadto w roku akademickim 2011-2012 prowadził specjalistyczne zajęcia z zakresu lingwistyki komputerowej dla doktorantów sąsiedniego Wydziału Anglistyki UAM, co niebagatelnie świadczy o szerokości horyzontów dydaktycznych Habilitanta (zajęcia te, to: „Application of computer tools for linguistic research" oraz „Translation and computers"). W latach 2011-2017 r. brał udział w procesie dyplomowania i sprawował opiekę naukową nad 15 magistrantami na rodzimym Wydziale Matematyki i Informatyki UAM, co świadczy o wysokiej ocenie naukowo-dydaktycznej Habilitanta, wystawionej przez najbliższe Mu środowisko naukowe. Tak wysoka ocena w tym zakresie zaowocowała także powierzeniem Habilitantowi funkcji promotora pomocniczego w przewodzie doktorskim prowadzonym w latach 2011-2017 na Wydziale Anglistyki UAM (co właściwie jest zewnętrznym potwierdzeniem wysokiej oceny w tym zakresie wystawionej przez najbliższe Mu środowisko naukowe).

Wysoką pozycję w szerokim środowisku badawczym wyznacza także i to, że powierzano habilitantowi recenzowanie tekstów przed publikacją w specjalistycznym czasopiśmie naukowym: „Studia Informatica".

Odmianą współpracy z otoczeniem prowadzonej przez Habilitanta jest Jego działalność popularyzatorska, którą także należy docenić. Przygotował 2 wykłady popularnonaukowe, w tym jeden w ramach szeroko zakrojonej akcji popularyzatorskiej UAM – podczas Festiwalu Nauki i Sztuki.

Należy zatem jednoznacznie stwierdzić, że i tu – w ramach działalności Habilitanta obejmującej pozostałe obszary aktywności naukowej, organizacyjnej, dydaktycznej i popularyzatorskiej – dr Graliński wykazuje się dużą aktywnością, przekraczającą niejednokrotnie oczekiwania kierowane pod adresem niesamodzielnych pracowników nauki i przewyższającą standardowe obciążenie i wymagania (głównie merytoryczne i dydaktyczne) nakładane na członków tej grupy pracowników nauki.

Konkluzja: Przegląd dorobku naukowego dra Filipa Gralińskiego z okresu po uzyskaniu stopnia doktora **w pełni pozwala ocenić ten dorobek jako osiągnięcie badawcze, wnoszące istotny wkład w rozwój uprawianej dyscypliny naukowej (językoznawstwa).** Habilitant prezentuje się bowiem jako doświadczony, a zarazem nowatorski badacz języka polskiego – a dokładniej: lingwista komputerowy, szczególnie kompetentny w zakresie lingwochronologii, korpusologii (w tym: korpusologii diachronicznej!) i wszelkich aspektów lingwistyki komputerowej – nowej dziedziny wiedzy powstającej na styku językoznawstwa i informatyki; dziedziny wiedzy o bardzo szerokim zastosowaniu w praktyce badawczej, jak i gospodarczej, że o najwyższych obecnie osiągnieciach tej dziedziny wiedzy – o sztucznej inteligencji czy o głębokim uczeniu, uczeniu maszynowym i przetwarzaniu języka naturalnego – można by tu wspomnieć. Habilitant wypracował tu już własne umiejętności warsztatowe (analityczne i syntetyczne) oraz koncepcje teoretyczno-metodologiczne. **Poddane ocenie publikacje i różnorodna działalność naukowa świadczą o istotnej aktywności naukowej Habilitanta**, co wraz z pozytywną opinią o działalności pozanaukowej (organizacyjnej, dydaktycznej, popularyzatorskiej itp.) uprawnia do stwierdzenia, że **dr Filip Graliński spełnia z naddatkiem ustawowe wymogi do uzyskania stopnia naukowego doktora habilitowanego.** Zgodnie więc z art. 16. Ust. 2 ustawy z dnia 14 marca 2003 o stopniach naukowych i tytule naukowym (z późniejszymi zmianami)

przedstawiona w niniejszym postępowaniu jako główne osiągnięcie badawcze monografia Against the Arrow of Time. Theory and Practice of Mining Massive Corpora of Polish Historical Texts for Linguistic and Historical Research jest istotnym osiągnięciem naukowym Habilitanta w okresie po uzyskaniu stopnia doktora, wnoszącym znaczny wkład w rozwój współczesnego językoznawstwa (w tym i zwłaszcza – polonistycznego) zorientowanego na interdyscyplinarne badania – tu: w zakresie lingwistyki komputerowej (w tym – przetwarzania języka naturalnego), lingwochronologii i korpusologii (w tym – diachronicznej).

Wniosek: Na podstawie przedstawionej pozytywnej oceny osiągnięcia naukowego i istotnej aktywności naukowej Habilitanta wyrażam pozytywną opinię w sprawie nadania **doktorowi Filipowi Gralińskiemu stopnia doktora habilitowanego w dziedzinie nauk humanistycznych w dyscyplinie językoznawstwo.**

Poznań, 25 maja 2020 r.

/dr hab. Michał Szczyszek/