Prof. dr hab. Wiesław Lubaszewski
Profesor emerytowany
Katedra Lingwistyki Komputerowej UJ
Katedra Informatyki AGH

Kraków, 14 czerwca 2020 r.

Ocena dorobku naukowego dra Filipa Gralińskiego dla potrzeb postępowania o nadanie stopnia doktora habilitowanego z dyscypliny językoznawstwo na zlecenie Centralnej Komisji ds. Stopni i Tytułów z dnia 4 marca 2020r.

Dorobek badawczy dra Gralińskiego lokuje się na styku informatyki i językoznawstwa. Warto więc przypomnieć, że związki językoznawstwa i informatyki mają długą tradycję i sięgają końca lat 40 XX wieku, w których podjęto pierwsze próby tłumaczenia maszynowego. W wyniku tych prób już na początku lat 50 XX wieku zdano sobie sprawę z tego, że aspekt technologiczny tłumaczenia jest nieporównanie prostszy od aspektu językoznawczego, gdyż to językoznawstwo powinno dostarczyć algorytmowi tłumaczącemu informacji koniecznej do poprawnej interpretacji tłumaczonej wypowiedzi. Świadomość tak sformułowanego zadania spowodowała intensywny rozwój badań językoznawczych, na czym skorzystała także informatyka. Większość informatyków styka się z gramatykami Chomsky'ego już w trakcie studiów, ale już nie wszyscy wiedzą o tym, że język XML jest pochodną prac nad standaryzacją struktury haseł w słowniku Oxfordzkim prowadzonych przez informatyków z *IBM*.

Dr Filip Graliński jest świadom tego, że jednym z podstawowych problemów, jaki dziś musi rozwiązywać językoznawstwo, chcąc dostarczyć algorytmom odpowiedniej informacji: fonetycznej, gramatycznej, semantycznej i pragmatycznej jest poszukiwanie modelu, który sprowadzałby abstrakcyjny językoznawczy opis formy i znaczenia jednostki słownika (leksemu) do takiego poziomu szczegółowości, by mógł z tego opisu korzystać algorytm przetwarzający napisy (ciągi liter). W praktyce oznacza to, że opis leksemu powinien informować o tym, że formy (do) *zamku* i *zamka* reprezentują dwie różne jednostki leksykalne a formy *zamek, zamki, zamków*, ... mogą w tekście reprezentować każdą z tych jednostek. Ponadto słownik powinien rejestrować fakt, że *baszta, wieża, fosa*, ..., *budować, zburzyć, oblegać* łączą się z jednostką, którą reprezentuje forma *zamku*, a *klucz, drzwi*, ..., *zamknąć, zepsuć, naprawić, zamontować* z tą reprezentowaną przez formę *zamka*. Natomiast *sprzedać, kupić*, .., łączą się z obu jednostkami. Rzecz jasna, budowa słownika o takim stopniu szczegółowości wymaga bardzo dużych zbiorów tekstów i algorytmów które przeprowadzą wstępną analizę łączliwości dostarczając przy okazji informacji statystycznych. Analiza wstępna powinna zostać zweryfikowana przez językoznawcę, który oceni pracę algorytmu, np. sprawdzając, czy *mieszkać* łączy się tylko z jedną jednostką leksykalną, a także czy *kupić* i *sprzedać* łączą się z obu jednostkami leksykalnymi, mimo tego, że sprzedaż zamków (warowni) jest zjawiskiem rzadkim, co odwzorują teksty. Zauważone błędy i ich interpretacja dostarczą danych do ulepszenia algorytmu. Korzyść z takiego postępowania odnosi także językoznawstwo. Dostęp do dużych zbiorów tekstów pozwala np. łatwiej weryfikować i zaktualizować zawartość słowników tradycyjnych. Językoznawcza analiza

wyników dostarczonych przez algorytm dostarczy także danych przydatnych dla semantyki leksykalnej i składni. Naszkicowana tu współzależność językoznawstwa i informatyki wyznacza obszar badań nazywanych leksykografią komputerową. Osiągnięcie główne dra Gralińskiego mieści się w tym właśnie obszarze.

Osiągnięcie główne

Monografia *Against the Arrow of Time. Theory and Practice of Mining Massive Corpora of Polish Historical Texts for Linguistic and Historic Research*, Wyd. UAM, Poznań 2019, składa się z 11 rozdziałów oraz spisów tekstów, rysunków, tabel, indeksu i bibliografii. Zadeklarowana jako osiągnięcie główne monografia stanowi podsumowanie podoktorskiego dorobku dra Gralińskiego, który pisze w przedmowie: "*This book may seem like a labirynth, as a number of topics are covered, from linguistics to computer science, folkloristcs, an library science. Depending on who the reader is, various reading strategiies can be applied.*" [s. 9]. Trzeba jednak na tę nieco kokieteryjną deklarację spojrzeć z dystansem i serio powiedzieć, że recenzowana monografia mogłaby mieć podtytuł: studia z leksykografii komputerowej. Autor bowiem przedstawia na wybranych przykładach niemal wszystkie najważniejsze aspekty tego obszaru badań, tj. dobór i przygotowanie tekstów, wybrane narzędzia automatycznej ekscerpcji danych leksykalnych oraz zastosowania tych narzędzi w badaniach wybranych zjawisk leksykalnych i dających się określić leksykalnie zjawisk współczesnego folkloru miejskiego. Do pełnego obrazu leksykografii komputerowej brak tylko rozważań nad problemem półautomatycznej konstrukcji słownika, jednak nie można z tego czynić zarzutu, gdyż byłby to temat wymagający obszernego, odrębnego opracowania. Materiał badań opisanych w monografii stanowią teksty prasowe z XIX i XX, więc monografia zajmuje się problemami leksykografii historycznej, jednak użyte w tytule określenie *Polish historical texts* trzeba uznać za przesadne.

Autor podzielił monografię na 4 części i podział ten trzeba uznać za uzasadniony, jednak części 2, 3 i 4 odnoszące się do różnych aspektów tego samego zagadnienia omówię łącznie.

Dobór i przygotowanie tekstów.

Problemowi danych językowych poświęca Autor pierwszą część monografii 1 *Textual mass*, składającą się z trzech rozdziałów: 1. *What is out there* (s. 19-32), 2. *Metadata* (s. 33-59), 3. *Texts* (s. 63-102). Dwa pierwsze rozdziały mają charakter techniczny, więc nie będę ich omawiał. Rozdział trzeci poświęcony tekstom ma znaczenie zasadnicze dla całej monografii, Autor bowiem opisuje nie tylko techniczne problemy związane z procesem dygitalizacji tekstów, ale też omawia problemy merytoryczne związane z kształtem gromadzonych danych tekstowych i uzasadnia decyzje, których skutki będą się przejawiać w prezentowanych językoznawcy wynikach algorytmów analitycznych omawianych w dalszych częściach monografii.

Podstawowy problem jaki musi rozwiązać badacz budujący korpus stanowią kryteria doboru tekstów, które powinny pozwolić na uzyskanie tzw. korpusu zrównoważonego, tj. takiego który reprezentuje wszystkie zjawiska językowe, i w którym wszystkie zjawiska językowe mają odpowiednią reprezentację statystyczną. Jest to postulat teoretyczny, który trudno zrealizować w praktyce. Dr Graliński świadomie rezygnuje z prób budowania korpusu zrównoważonego, gdyż uważa, że statystyczna procedura równoważenia może usunąć z

korpusu zjawiska względnie rzadkie. Buduje zatem korpus tekstów prasowych obejmujący dwa stulecia stosując kryteria praktyczne. Nie mam wątpliwości, że Habilitant dokonał trafnego wyboru, bowiem przyszłe porównanie języka prasy dziewiętnastowiecznej z lepiej znanym językiem prozaików tego okresu będzie cenne z językoznawczego punktu widzenia.

Problem drugi to poprawność zdygitalizowanych tekstów. Wiadomo, że optyczny czytnik pisma (OCR) to rządzenie, którego działanie jest zależne od jakości układu optycznego, który można kalibrować dostosowując urządzenie do rodzaju papieru, kształtu czcionki itd. Jednak kalibracja nie rozwiązuje problemów wynikających z niewidocznego dla oka uszkodzenia liter. W takiej sytuacji algorytm rozpoznawania liter może nie umieć jednoznacznie rozstrzygnąć, jakiemu ze znanych algorytmowi wzorców liter odpowiada zeskanowany z tekstu kształt. Rozstrzygnięcie na podstawie informacji o prawdopodobieństwie wystąpienia konkretnej litery może spowodować błąd. Wiemy z monografii, że algorytm o opisanym kształcie popełnia wiele błędów, gdy zastosujemy go do tekstów dawnych. Niewątpliwie, potrzebny jest algorytm wspomagający, który na podstawie kontekstu, tj. liter otaczających niejednoznaczny kształt zaproponuje poprawne rozwiązanie. Jednak nie wystarczy tu stosowany w informatyce kontekst dwu lub trzyliterowy, bowiem nie rozstrzyga on np. przypadku *sąd* : *sad*. Zgadzam się z dr Gralińskim, że poszukiwanie metod zapobiegania błędom OCR jest problemem językoznawczym. Autor nie proponuje własnego algorytmu pomocniczego, lecz przedstawia dwie metody automatycznego wykrywania błędów w zdygitalizownych tekstach, co może dostarczyć danych twórcom algorytmu.

Problem trzeci to uwspółcześnianie tekstów dawnych, co w terminologii stosowanej przez Autora nazywa się normalizacją tekstów historycznych. Problem uwspółcześnienia jest dobrze znany wszystkim wydawcom tekstów dawnych, którzy z reguły unowocześniali pisownię, często podawali współczesną wersję nazw własnych, rzadko jednak i bardzo ostrożnie unowocześniali gramatykę i leksykę, chcąc zachować oryginalne walory stylu konkretnego autora. Dr Graliński chciał tak zmodernizować teksty z XIX i XX wieku, by można do ich analizy użyć narzędzi zbudowanych dla współczesnej polszczyzny, czyli słowników fleksyjnych i analizatorów morfologicznych, tj. algorytmów interpretujących formę fleksyjną występującą w tekście i wskazujących leksem (jednostkę słownika), który dana forma reprezentuje. Autor nie ma wątpliwości, że w pierwszym rzędzie należy uwspółcześnić pisownię tekstów dawnych. W wyniku prac nad automatyczną konwersją ortograficzną powstał wartościowy, empirycznie zweryfikowany zbiór reguł konwersji ortografii stosowanej w XIX i XX w do postaci współczesnej – zob. s. 94-96. Trzeba też podkreślić, że dr Graliński podjął trafną decyzję językoznawczą, rezygnując z dalszych możliwych etapów modernizacji tekstów.

Automatyczna ekscerpcja danych z korpusu tekstów

W klasycznej leksykografii ekscerpcji danych z tekstów dokonywali językoznawcy, zapisując ekscerpty na fiszkach, często wraz z interpretacją wyekscerpowanego zjawiska. Zazwyczaj dokonujący ekscerpcji językoznawca nie ograniczał się do poszukiwania konkretnego zjawiska, np. form dopełniacza liczby mnogiej rzeczowników nijakich, lecz ekscerpował wszystko to, co uznał za istotne, zwłaszcza zjawiska rzadkie lub nietypowe. W wyniku takiego postępowania ekscerpcja ręczna dawała wynik o wysokiej jakości, czego dowodzą np.

przechowywane w kartotece *Słownika staropolskiego* ekscerpty z rękopisów wykonane przez wielkich językoznawców. Ekscerpcja automatyczna stawia językoznawcę w innej sytuacji. Nie czyta on bowiem tekstów lecz ekscerpty wyszukane przez algorytm. To zaś rodzi pytania o jakość ekscerpcji, zwłaszcza o to, czy algorytm potrafi ekscerpować zjawiska rzadkie i nietypowe.
Różnym aspektom automatycznej ekscerpcji danych z tekstów dawnych poświęca dr Graliński pozostałe części monografii.

Część 2 *(Re)searching* składająca się z rozdziałów 4. *Searching for words* (s. 103-136) i 5. *From search into reresarch* (s. 137-150) jest poświęcona wyszukiwaniu zjawisk językowych w tekstach. Wyszukiwanie opisane w tej części to tzw. wyszukiwanie boolowskie (boolean search) polegające na tym, że algorytm wyszukujący dostaje na wejściu wzorzec zbudowany z ciągów liter i zwraca wszystkie identyczne ciągi liter występujące w przeszukiwanych tekstach. Ciągi liter tworzące wzorzec mogą być morfemami, formami fleksyjnymi, leksemami, wyrażeniami, zdaniami itd. Algorytm wyszukujący może na życzenie podawać konteksty w jakich występuje poszukiwany wzorzec a także liczbę wystąpień konkretnego wzorca. Bez wątpienia, algorytm tego typu jest niezwykle przydatny w badaniach językoznawczych, jednak nakłada on na językoznawcę obowiązek wyobrażenia sobie zjawisk rzadkich i nietypowych, gdyż, jak powiedziałem, wyszukiwanie boolowskie wyszukuje tylko to, co zostało zdefiniowane we wzorcu. Przy wyszukiwaniu zjawisk rzadkich i nietypowych pomocne może się okazać wyszukiwanie przybliżone, które może przeprowadzić algorytm statystyczny, np. taki jak algorytm Churcha i Hanksa (*Word Association Norms, Mutual Information, and Lexicography*, Computational Linguistics, June 2002), który sam wyszukuje w tekstach powiązania leksemu wejściowego z innymi leksemami, zwracając listę uporządkowaną według siły skojarzenia, a więc listę podobną do listy uzyskanej w wyniku eksperymentu swobodnych skojarzeń słownych, który zainspirował twórców algorytmu. Skonstruowana w ten sposób lista może zawierać powiązania rzadkie i nietypowe. Wymieniany w monografii system „Odkrywka", na który składają się algorytmy opisane w pracy zawiera inny, bardziej skomplikowany algorytm statystyczny omówiony w dalszych częściach pracy, jednak sądzę, że warto pamiętać o algorytmie klasycznym, gdyż jest przydatny dla językoznawcy.

Każde wyszukiwanie jest w równym stopniu zależne od przygotowania tekstów, na których operują algorytmy. Problem wykracza zdecydowanie poza konwersję ortograficzną. Dr Graliński rzetelnie omawia zagadnienie wskazując przykłady błędów, wynikających z niewłaściwego przygotowania tekstów. Jednak eliminacja błędów wynikających np. z homografii form fleksyjnych, (np. *mam* reprezentuje *mieć* lub *mama*) to poważne zadanie językoznawcze – mam na myśli opracowanie reguł rozstrzygania, a nie ręczny opis tekstów. Opracowanie reguł rozstrzygania homografii dla dawniejszych warstw polszczyzny to zadanie pilne, co pokazuje praca dra Gralińskiego.

Kolejna część monografii, tj. 3 *Modelling* opisuje statystyczne podstawy modelowania zjawisk językowych w rozdziale 6. *Temporal language models* (s. 151-170), pokazując zastosowania modeli statystycznych w rozdziale 7. *Temporal text classification* (s.171-184). Z językoznawczego punktu widzenia istotny jest rozdział 8. *Word embedding for diachrony*

(s. 185-224), który przedstawia ciekawą propozycję statystycznego modelowania łączliwości semantycznej leksemu na podstawie korpusu tekstów. Algorytmy statystyczne operują na liczbach, więc algorytm modelujący łączliwość semantyczną dwu leksemów może mieć na wejściu np. trzy liczby, tj. częstość występowania leksemu *a* i leksemu *b* niezależnie oraz częstość wspólnego wystąpienia *a* i *b* w tzw. oknie (otoczeniu leksykalnym) liczącym np. 5 wyrazów. Zależność pomiędzy liczbami można także przedstawić jako liczbę, której wartość wylicza się za pomocą wzoru przyjętego dla modelowania konkretnego zjawiska i która w algorytmie Churcha i Hanksa wyraża bliskość (siłę skojarzenia) *a* i *b*. Testowany w monografii model *Word2vec* (Word to Vector) oblicza zależność między *a* i *b* w sposób znacznie bardziej skomplikowany (s.191-193), a liczba charakteryzująca powiązanie *a* i *b* jest wpisywana do wektora leksemu definiowanego. Algorytm Word2vec najpierw „uczy się", tj. na podstawie wybranych przez człowieka przykładów buduje charakterystykę liczbową powiązań pozytywnych, np. *a* i *b, a* i *c,* itd. Następnie algorytm analizuje zbiór tekstów rozbudowując i modyfikując zbiór liczb charakteryzujących powiązania konkretnego leksemu. W rezultacie wektor leksemu definiowanego to zbiór liczb uzyskanych dzięki obliczeniom uwzględniającym leksemy najczęściej współwystępujące z definiowanym. Zbudowane w ten sposób charakterystyki leksemów można porównywać, bowiem leksemy mające podobną łączliwość w tekście powinny mieć podobną charakterystykę liczbową (wektory). Jednakże tabela 8.3 na s. 192, pokazująca relacje podobieństwa dla 5 leksemów budzi odczucia mieszane. Wektor nazwy *Polska* ma jako najbardziej podobne wektory ciągów liter *Połska, Folska, Pełska* itd, które powstały w wyniku błędu optycznego czytnika tekstów. *Warszawa* oprócz błędów czytnika ma na liście podobieństwa nazwy niektórych miast. Podobnie zachowują się nazwy *Niemcy* i *Berlin*. Zdecydowanie lepiej wygląda homograficzna forma *kaszle* - tu na liście podobieństwa dominują formy leksemów nazywających różne objawy chorobowe: *kaszla, wymiotuje, kurcze, chrypke, kicha* itd. Jednak dr Graliński nie traci entuzjazmu badawczego i w pełni się zgadzam z taką postawą. To tylko eksperyment, do którego powinno się włączyć językoznawstwo i być może zdoła pomóc informatyce. Model Word2vec jest siecią neuronową i można stosunkowo łatwo zmieniać jej architekturę. Skoro więc wiemy, że łączliwością rzeczownika, przymiotnika i czasownika rządzą różne reguły, np. *biały* może się odnosić do każdego niemal rzeczownika, a *szczeka* w zasadzie tylko *pies*, to warto zrezygnować z podejścia *one size fits all* i poszukać eksperymentalnie architektur sieci odpowiednich dla głównych części mowy. Można też podzielić opisany w monografii korpus na dwa odrębne. Pierwszy zawierający teksty wydane do r. 1918 a drugi po 1918 r., by następnie zbudować odrębne charakterystyki dla *Warszawy* i innych nazw. Ciekawe, czy będą się różnić?

Wektorowe reprezentacje leksemów można także dodawać i odejmować. Wspomniane operacje służą Autorowi do poszukiwania analogii leksykalnych, np. *aeroplan : samolot*, a także do oceny kwalifikatorów historycznych w słowniku Doroszewskiego.

Ostania część monografii 4 *Applications,* składająca się z rozdziałów 9. *Lexical ephemera* (s. 225-240), 10. *Traps of cultutronix* (s. 241-256) i 11. *Folkloristics 2.0* (s. 257-288), jest poświęcona automatycznej ekstrakcji leksyki reprezentatywnej dla zmian historycznych i

społecznych, a także leksyki identyfikującej stereotyp kulturowy. Nie będę ich jednak omawiał szerzej, gdyż opisują zastosowania omówionych wcześniej narzędzi i metod.

Podsumowując, mogę powiedzieć, że monografia przedstawia oryginalne i wartościowe badania z zakresu leksykografii komputerowej, które warto kontynuować. Uwzględniając opisane na wstępie recenzji usytuowanie leksykografii komputerowej mogę z przekonaniem stwierdzić, że osiągnięcie główne dra Filipa Gralińskiego spełnia wymogi merytoryczne stawiane w postępowaniu habilitacyjnym z językoznawstwa.

Pozostały dorobek

Dr Filip Graliński opublikował ponad 30 artykułów w czasopismach i tomach zbiorowych. W tej części dorobku liczbowo przeważa problematyka leksykograficzno-komputerowa. Wśród tekstów przedstawiających badania nie włączone do monografii wyróżniają się indeksowana autorska praca nad rozpoznawaniem idiomów za pomocą wyrażeń wprowadzających: *Mining the Web for Idiomatic Expressions Using Metalinguistic Markers* LNCS, Springer, Berlin, 2010 oraz systematyczny opis systemu leksykograficznego Odkrywka: *System Odkrywka jako innowacyjne narzędzie do badania polskiej leksyki potocznej*, wspólnie z D. Dzienisiewiczem i K. Świetlikiem, w: A. Piotrowicz, M. Witaszek-Samborska, K. Skibiński red. Kultura komunikacji potocznej, WNI, 2018.

Problematyka leksykograficzno-komputerowa to także temat większości referatów wygłoszonych przez dra Gralińskiego na konferencjach międzynarodowych.

Wśród publikacji informatycznych można wyróżnić np. badania poświęcone automatycznej analizie nieczytelnych dla człowieka wyników algorytmów machine learning, operujących na danych tekstowych, co ma ułatwić analizę działania algorytmu i danych na których operuje (*GEval: Tool for Debugging NLP Datasets and Models*, wspólnie z A. Wróblewską i T Stanisławkiem, ACL, Florencja, 2019) oraz budowa środowiska pozwalającego na porównawczą analizę algorytmów i zbiorów danych (*Gonito.net – Open Platform for Research Competition, Cooperation and Reproductability*, wspólnie z R. Jaworskim, Ł. Borchmannem i P. Wierchoniem, w; P. Sojka i inni red, LNCS, Heidelberg 2012).

Podsumowując mogę stwierdzić, że także ta część dorobku naukowego pozwala doktorowi Filipowi Gralińskiemu na ubieganie się o stopień doktora habilitowanego z językoznawstwa.

Konkluzja

Przeprowadzona w recenzji ocena całości dorobku dra Filipa Gralińskiego jest pozytywna, a dorobek ten spełnia wymogi merytoryczne stawiane kandydatom ubiegającym się o stopień doktora habilitowanego z językoznawstwa.