Poznańskie Studia Polonistyczne
Seria Językoznawcza
vol. 28 (48), nr 1
DOI: 10.14746/pspsj.2021.28.1.22

Yury Fedorushkov, *Adaptacja warsztatu leksykograficznego do automatyzacji rozpoznawania frazemów w systemie uczącym się (na przykładzie konstrukcji werbo-nominalnych języka rosyjskiego)*, Wydawnictwo Naukowe UAM, Poznań 2019, ss. 292

Recenzowana książka Yurego Fedorushkova, *Adaptacja warsztatu leksykograficznego do automatyzacji rozpoznawania frazemów w systemie uczącym się (na przykładzie konstrukcji werbo-nominalnych języka rosyjskiego)*, jest pozycją niezwykle interesującą z naukowego, materiałowego i poznawczego punktu widzenia. Powinna stać się obowiązkową lekturą na kierunkach językoznawczych i informatycznych; powinna być czytana zwłaszcza przez badaczy oraz studentów specjalizujących się w szybko rozwijającej się obecnie dziedzinie wiedzy nazywanej lingwistyką komputerową, której szczytowym osiągnięciem są badania w zakresie sztucznej inteligencji. Niewykluczone, że i przedstawiciele nauk prawnych, a w szczególności kryminolodzy, mogliby znaleźć inspiracje dla swojej pracy badawczej.

W książce omówiono wyniki badań eksperymentalnych dotyczących jednego z aspektów językoznawstwa – zbadano frazemy języka rosyjskiego, a dokładniej: automatyczne, tj. z wykorzystaniem stosownego algorytmu informatycznego, sposoby wykrywania frazemów języka rosyjskiego w tekstach (oczywiście – rosyjskojęzycznych). To zagadnienie stanowi kluczowy naukowo aspekt recenzowanej pracy. Fedorushkov przedstawił bowiem bardzo skomplikowaną metodologicznie i technologicznie procedurę wyłuskiwania narzędziami informatycznymi (wspomnianym algorytmem, całą procedurą ekscerpcyjną) frazemów z korpusu tekstów. Co ważne – procedura ta została nie tylko opracowana teoretycznie, ale także sprawdzona w praktyce (kilkuetapowo) na wielu korpusach elektronicznych, w tym – na ogromnym materiale weryfikacyjnym. Materiał ten to korpus języka rosyjskiego liczący nie miliony, lecz miliardy segmentów-jednostek! Korpus – zasób o nazwie Common Crawl – został stworzony na potrzeby badań, testowania, tj. „trenowania”

specjalistycznego narzędzia informatycznego i całej procedury stosowanej do ekscerpcji materiału frazematycznego. Trzeba dodać, że – według mojej wiedzy korpusologicznej – większe korpusy są zgromadzone tylko na Uniwersytecie Stanforda dla współczesnego języka angielskiego. Czytelnik – całe środowisko badaczy języka, lingwiści komputerowi, specjaliści w zakresie sztucznej inteligencji – uzyskał zatem nie tylko niezwykle precyzyjny opis teoretyczny samego narzędzia i procedury ekscerpowania frazemów, ale także sprawdzone, tj. skontrolowane w praktyce wyniki samej ekscerpcji frazemów z nieotagowanych tekstów języka rosyjskiego. Wynik badawczy jest zatem pozytywnie zweryfikowany przez bardzo dobrze przygotowany, logicznie uzasadniony i wewnętrznie niesprzeczny eksperyment lingwistyczno-komputerowy, który autor tu zastosował.

Teoretyczną koncepcję frazemów Fedorushkov zaczerpnął z prac polskiego badacza tej problematyki – Wojciecha Chlebdy. I zaadaptował ją twórczo i z wielkim pożytkiem dla swoich badań. „Przełożył" bowiem tę koncepcję na język „zrozumiały" dla komputerów, dla narzędzi matematyczno-informatycznych z zakresu lingwistyki komputerowej. Zrobił to, odwołując się do pojęcia n-gramów (bigramów, trigramów itp.), co dało niezwykle precyzyjne wyniki badawcze, uzyskane w zastosowanym eksperymencie, tj. przez testowane w jego trakcie narzędzie informatyczne do ekscerpcji frazemów. Wyniki mają dużą wartość zarówno naukową, jak i w zakresie stosowalności praktycznej. Dzięki pracy Fedorushkova można z dużą dokładnością orzec dziś, czy w strumieniu tekstów pojawiają się frazemy (i neofrazemy), czy też nie. Oczywiście skuteczność narzędzia jest teraz „ograniczona" do jednego języka (rosyjskiego), ale to „ograniczenie" łatwo będzie zniwelować w dalszych pracach, badaniach i eksperymentach autora, do czego go zachęcam! Rezultaty badawcze (i zastosowanie praktyczne) wypracowane tu dla języka rosyjskiego mogą stanowić podstawę do stworzenia podobnych narzędzi (lub zastosowania tego narzędzia) dla innych języków, np. polskiego, angielskiego, niemieckiego.

To wszystko, co powyżej ująłem w planie naukowym, metodologiczno-technologicznym i materiałowym, jest niezwykle ważne i inspirujące w planie poznawczym. Otóż recenzowana książka otwiera wiele daleko sięgających planów naukowych i zaawansowanych badań w różnych obszarach badawczych. O kilku z nich już wspomniałem. Inne to np. możliwość wykorzystania opracowanego przez Fedorushkova narzędzia do określania autorstwa tekstów anonimowych lub o niepewnym autorstwie – a piszę to jako czynny biegły sądowy z zakresu językoznawstwa, nierzadko spotykający się ze sprawami sądowymi, w których takie zbadanie tekstu jest bardzo ważne czy wręcz kluczowe. Niemniej trzeba dodać, że sam autor ten wątek w książce tylko sygnalizuje i go –

niestety – nie rozwija. To może budzić obawy, czy narzędzie to zostało już zweryfikowane również pod tym kątem. Ponadto dzięki książce Fedorushkova i opisywanym w niej narzędziu można pokusić się o przewidywanie dalekosiężnych konsekwencji w badaniach literaturoznawczych, historycznych, kulturoznawczych, etnograficznych. Recenzowana książka wnosi także niemały wkład – teoretyczny, metodologiczno-technologiczny, czy w końcu praktyczny – do nauk informatycznych, zwłaszcza do wspomnianych lingwistyki stosowanej czy sztucznej inteligencji (stojących na pograniczu językoznawstwa i informatyki). I oczywiście – sama lingwistyka też zyskuje niezwykle precyzyjne narzędzie badawcze oraz dokładnie zweryfikowane rezultaty naukowe – w zakresie ekscerpcji jednostek wielosegmentowych (frazemów, frazeologizmów, przysłów itp.) i możliwości badania tak skonstruowanego niemałego zbioru tego rodzaju jednostek języka.

Uważam zatem, co piszę z pełną odpowiedzialnością, że recenzowana książka to publikacja dokumentująca oryginalne i twórcze osiągnięcie naukowe i mająca istotny wpływ na stan wiedzy i kierunki dalszych badań – w kilku wymienionych wyżej potencjalnych obszarach, dziedzinach, dyscyplinach. W przyszłości może się okazać, że tych dziedzin będzie więcej, dziś trudno o tym wyrokować.

Michał Szczyszek