Prof. dr hab. B. Krzysztof Bogacki

Uniwersytet Warszawski

Recenzja rozprawy doktorskiej mgr Agnieszki Palion-Musioł

pt. **Desambiguacion semantica automatica de los verbos seleccionados de movimiento *salir* y *entrar* en el enfoque orientado a objetos/Automatyczne odwieloznacznienie semantyczne wybranych czasowników ruchu *salir* i *entrar* w ujęciu zorientowanym obiektowo**

Pani mgr Agnieszka Palion-Musioł przedłożyła rozprawę pod wskazanym wyżej tytułem wykonaną pod kierunkiem prof. dra hab. Wiesława Banysia w przewodzie doktorskim otwartym na Wydziale Filologicznym Uniwersytetu Śląskiego. Celem badań, jaki przed sobą postawiła jest szczegółowe wyjaśnienie mechanizmów dezambiguizacji czasowników hiszpańskich, co jest pokazane na przykładzie dwóch wysoce polisemicznych jednostek leksykalnych oznaczających przemieszczanie: *salir* oraz *entrar*. Wybór, jakiego dokonała jest bardzo szczęśliwy i bez wątpienia spełnia wymogi stawiane przez odpowiednie przepisy przed dysertacjami doktorskimi. Zagadnienie dezambiguizacji należy do kluczowych nie tylko z punktu widzenia teoretycznego, ale także praktycznego. W rzeczy samej, bez sprawnie funkcjonującego rozwiązania problemu odwieloznaczniania nie ma mowy o postępie w tłumaczeniu maszynowym, bowiem program informatyczny musi zawierać wskazówkę który odpowiednik danego wyrazu w języku źródłowym należy wybrać ze słownika transferowego dla języka docelowego (w tym wypadku jest nim język polski). Metod usuwania wieloznaczności jest kilka, w tym proponowana i stosowana z powodzeniem w śląskim ośrodku naukowym metoda syntaktyczno-semantyczna wysuwająca na pierwszy plan zorientowanie obiektowe zaproponowana przez Promotora Doktorantki, prof. dra hab. Wiesława Banysia. Według tej metody wykonano na Uniwersytecie Śląskim szereg prac doktorskich. Poczynając od rozpraw S. Szramek-Karcz, B. Śmigielskiej z 2006 roku, kończąc na niedawno obronionej dysertacji M.

Hrabi o wybranych czasownikach ruchu mieliśmy serię prac, które udowodniły z jednej strony uniwersalność metody gdy chodzi o brane pod uwagę języki (badania dotyczyły nie tylko języków romańskich w zestawieniu z polskim – francuski, hiszpański ale także angielski, por. Drzazga A., 2012, **The disambiguation of the English verbs *send* and *open* – a study based on the object oriented method**). Z drugiej strony uwzględniane były różnorodne struktury językowe. Większość badań dotyczyła czasowników, ale pojawiła się też w 2008 roku praca A. Chrupały pt. **Les noms composés avec *femme* en français: une étude de leur degré de figement en vue d'un traitement automatique**. Oprócz poważniejszych prac na stopień, powstały także mniej obszerne studia. W ostatnich latach sama Doktorantka opublikowała dwa artykuły poświęcone tej tematyce: jeden samodzielnie w **Linguistica Silesiana**, drugi we współpracy z A. Żłobińską-Nowak w **Neophilologica**. Można powiedzieć, że daleko nam jeszcze do wyczerpania tematu. Problemem jest raczej znalezienie sposobu na przyspieszenie analiz, które pokryć powinny całość słownictwa składającego się z przeważającej większości z leksemów polisemicznych.

Przesłany mi do oceny tekst dysertacji doktorskiej jest jak na rozprawę tego typu bardzo obszerny, liczy bowiem 715 stron. Został zredagowany po hiszpańsku i rozpada się na 3 części (w ten sposób będę tłumaczył używany przez Autorkę termin „capítulos"). Całość otwiera wstęp (str. 6-45) przedstawiający ogólną problematykę tłumaczenia maszynowego. Część pierwsza (**Fundamentos teoricos lexicograficos**, str. 46-191) przedstawia teorię, w ramach której zostaną zanalizowane dwa hiszpańskie czasowniki ruchu i użyte w tym celu narzędzia. Część druga (**Analisis sintactico-semantica de los verbos de desplazamiento *salir* y *entrar* – parte practica**, str. 192-690) zawiera prezentację materiału językowego i w sposób naturalny dzieli się na dwie partie. Autorka przytacza za istniejącymi słownikami dane dotyczące semantyki i składni omawianych leksemów oraz proponuje schematy składniowo-semantyczne, których elementy składowe są następnie bardzo bogato ilustrowane. Część trzecia wreszcie to wnioski (**Conclusiones** str. 691-694), po niej zaś czytelnik znajdzie bibliografię (str. 695-709) i trzy streszczenia: w języku hiszpańskim, polskim i angielskim.

Nie bez kozery wskazałem strony każdej części, zabieg ten pozwala bowiem zdać sobie sprawę, że o ile sam u k ł a d elementów treści należy uznać za poprawny (najpierw zagadnienia teoretyczne, potem zaś prezentacja rezultatów materiałowych), o tyle ich podział na części nie został dokonany w sposób optymalny. Na jednym poziomie zostały bowiem umieszczone części (capítulos) o znacznie zróżnicowanych rozmiarach. Obok 3-stronicowego **Capitulo tercero** z wnioskami, znalazł się **capitulo secundo** liczący blisko 400 stron! Wolałbym podzielić ten ostatni na dwa: jeden dla *salir*, drugi dla *entrar*, a wnioskom nie nadawać statusu rozdziału.

W bibliografii znalazłem usterkę wynikłą zapewne z bezkrytycznego zawierzenia procesorowi tekstów w procesie sortowania. Spis literatury zawiera pozycje przywołane w tekście, co jest zgodne z regułami sztuki. Został on podzielony na kilka części: najpierw znajdujemy listę studiów, artykułów i monografii (w sumie 185 pozycji, dobrze dobranych i dobrze wykorzystanych), dalej słowniki (10 pozycji), źródła internetowe (13 pozycji) i dwie bazy danych. Otóż wydaje się, że ograniczona do adresu pozycja 96 z pierwszej listy - (http://www.tdx.cat/bitstream/handle/10803/83983/tg1de1.pdf;jsessionid=84B8CE54CB60CF220E5D57629E5D1A6F.tdx2?sequence=1) powinna zostać połączona z pozycją 86 odsyłającą do pracy doktorskiej obronionej na Universidad Autónoma de Barcelona, (Gyska T., 2011, tesis doctoral disponible en la pagina web). Należałoby też oczekiwać wskazania co najmniej tytułu dysertacji; tego niestety zabrakło.

Każda bibliografia jest odzwierciedleniem treści zawartych w pracy. Otóż czytelnik interesujący się na co dzień tłumaczeniem maszynowym z jednej strony z uznaniem musi przyznać, że Doktorantka panuje nad materiałem z tego zakresu, porusza się po nim z dużą swobodą i proponuje bardzo klarowny przegląd zastanego stanu rzeczy. Z drugiej strony jednak dostrzeże fakt, że swoje uwagi na temat obserwowanych tu trendów Autorka ogranicza w czasie. Trudno czynić z tego poważny zarzut, tak się bowiem składa, że ewolucja w tym zakresie postępuje z szybkością piorunującą, na dodatek zaś ostatnie nowości, które moim zdaniem, mają szansę stanowić znaczący krok na przód, stały się dostępne publicznie zaledwie kilka tygodni temu. Myślę tu o systemie opracowanym wspólnie przez Uniwersytet Harvarda i bardzo doświadczoną firmę Systran nazwanym Open-source Neural

Machine Translation System (http://opennmt.net/) . System ten, udostępniony o dziwo z bardzo oszczędnym kodem źródłowym ograniczonym do ok. 4.000 linii (sic!), oparty jest na idei sieci neuronowych. Pojawia się on po systemach bazujących na statystyce, z których w pewnym sensie się wywodzi, wykorzystuje także sztuczną inteligencję i w sumie zdaje się stanowić bardzo obiecującą alternatywę dla dotychczasowych rozwiązań. Czy więc w kontekście oceny pracy doktorskiej, pominięcie milczeniem informacji o systemie opartym na sieciach neuronowych stanowi czynnik umniejszający jej wartość ? Nie sądzę, bowiem jest jasne, że w informatycznej architekturze neuronowej da się z powodzeniem wykorzystać opis zaproponowanych przez Doktorantkę faktów ściśle językowych, podobnie jak było to możliwe w przypadku systemów statystycznych, czy jeszcze starszych systemów transferowych opartych na regułach.

Opis ten wykonany jest w ramach teoretycznych uwzględniających kilka elementów: model leksykograficzny Gastona Grossa, strukturę predykatowo–argumentową Stanisława Karolaka, teorię Sens-Tekst Igora Mielczuka, słownik w ujęciu frames i scripts pomysłu James'a Pustejovskiego i Branimira Boguraeva i wreszcie rozwiązania Xaviera Blanco i jego współpracowników z Barcelony.

Autorka materiał do opisu wybierała z jednej strony z dobrej jakości, renomowanych słowników, z drugiej zaś wspierała się danymi wydobytymi z dwóch elektronicznych baz danych : Marka Daviesa **Corpus del español: 100 million words, 1200s-1900s** (http://www.corpusdelespanol.org/) i Real Academia Española: **Banco de datos CREA** (en linea), **Corpus de referencia del español actual** http://www.rae.es. Wsparcie korpusowe ułatwiło przygotowanie schematów syntaktyczno-semantycznych. Punktem odniesienia dla obu leksemów hiszpańskich jest język polski, dlatego też nie zabrakło źródeł polskich (Perlin O., Perlin J., 2001, **Gran diccionario español-polaco**, Warszawa, Wiedza Powszechna i **Słownik języka polskiego**, 1993, Warszawa, Wydawnictwo Naukowe PWN). Opis uwzględnia kluczowe dla koncepcji G. Grossa pojęcie klas obiektowych. Co za tym idzie wskazane są czasowniki operatory i typowe dla danej klasy atrybuty. Wyodrębnione w ten sposób elementy są zebrane w tabelach przeglądowych uwzględniających ogromny materiał.

Jak wspomniałem wyżej praca jest wyjątkowo obszerna i ogromny udział mają w tym tabele podające długie listy jednostek leksykalnych nadających się do wypełnienia takiej czy innej pozycji argumentowej. Otóż można by je skrócić. Pierwszy sposób, to przejrzenie pod kątem wystąpienia ewentualnych dubletów, jak to ma miejsce na stronie 365, gdzie w tabeli CO 63 hasło „Internet" występuje dwukrotnie. Drugi sposób mógłby przynieść znacznie bardziej spektakularną kompresję. Warto moim zdaniem zastanowić się nad zasadnością wyliczania wyrazów, których końca nie widać a które, nadto, nie dają gwarancji kompletności. Oto przykład. Na stronach 358-361 znajdujemy wykaz 129 jednostek, które mogą funkcjonować jako CO 61. Otrzymały one etykietkę : **planta floreciente; flor czyli roślina kwitnąca; kwiat.** Otóż bez najmniejszego trudu czytelnik dotrze do internetowego **Glosario botánico** umieszczonego na stronie pod adresem (http://www.botanical-online.com/spanishglossary3.htm), i znajdzie w nim informację, że sama rodzina orchidei liczy ponad 25.000 roślin (każda musi mieć nazwę!), natomiast mniej liczna zwana papilionacea obejmuje 12.000 gatunków. Nie sposób wymienić je wszystkie, znają je zapewne i posługują się na co dzień botanicy. Podobne uwagi można zrobić odnośnie innych klas: rzeczowników oznaczających miejsca, nazwy zawodów itd. Zwykłemu śmiertelnikowi wystarczyłoby chyba podanie precyzyjnie sformułowanej etykietki (**planta floreciente, flor / roślina kwitnąca, kwiat**), tak jak zostało to zrobione w nagłówku tabeli i ograniczenie się do znacznie mniejszej liczby przykładów. Warto w tym miejscu zwrócić uwagę, że w takiej sytuacji ogromnego znaczenia nabiera precyzyjne rozgraniczenie klas, co nie zawsze jest proste. I tak Autorka w tabeli CO 30 (str. 324) umieszcza argumenty o charakterystyce **„miejsce lub szeroka droga, do której schodzą się różne ulice: nazwa"**, takie jak **GranVía** czy **rambla.** Z całą pewnością są to „rzeczowniki oznaczające miejsce", co prawda bardzo specyficzne. Otóż taką charakterystykę dała Autorka rzeczownikom wymienionym w tabeli CO 57 (str. 352-357), co prowadziłoby do wniosku, że CO 30 jest podzbiorem CO 57. Czy wobec tego rzeczowniki z CO 30 mogą się pojawić tam, gdzie figurują te z CO 57?

Zaletą pracy jest wyjątkowo drobiazgowe opracowanie schematów składniowo-semantycznych, które dzięki opozycjom występującym między różnymi występującymi w nich parametrami dają jednoznaczną charakterystykę wybranego

leksemu w języku źródłowym, to zaś z kolei pozwala jednoznacznie przyporządkować odpowiedni leksem w języku docelowym. Sieć tych opozycji jest zaiste imponująca, jej utworzenie zostało dokonane z ogromną precyzją, co wymagało benedyktyńskiej pracy. W efekcie otrzymaliśmy z tekst łączący w sobie oczywisty walor teoretyczny z dużą wartością praktyczną.

Wspomniane powyżej fakty prowadzą do wniosku, że przedłożona mi do oceny praca doktorska mgr Agnieszki Palion-Musioł pt. „**Desambiguacion semantica automatica de los verbos seleccionados de movimiento *salir* y *entrar* en el enfoque orientado a objetos/Automatyczne odwieloznacznienie semantyczne wybranych czasownikow ruchu *salir* i *entrar* w ujęciu zorientowanym obiektowo**"

spełnia wymagania stawiane rozprawom doktorskim.

W szczególności stanowi ona oryginalne rozwiązanie ciekawego problemu naukowego mającego nadto poważne implikacje praktyczne, jakimi jest dezambiguizacja leksemów wieloznacznych i stanowi dowód ogólnej wiedzy teoretycznej Doktorantki w zakresie językoznawstwa oraz umiejętności samodzielnego prowadzenia pracy naukowej. Biorąc zatem pod uwagę ustawę z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (por. Dz. U. nr 65, poz. 595, art. 13. 1 z późn. zm.) wnoszę o dopuszczenie jej Autorki do dalszych etapów przewodu doktorskiego.

Prof. dr hab. B. Krzysztof Bogacki
Uniwersytet Warszawski

Warszawa, dnia 15 lutego 2017 roku