Justyna Małek
Uniwersytet Śląski w Katowicach

Forum Lingwistyczne
nr 5, 2018

Sprawozdanie z seminarium naukowego „Przestrzenie humanistyki cyfrowej – korpusy diachroniczne polszczyzny" (Katowice, 24–25 kwietnia 2017)

W dniach 24–25 kwietnia 2017 roku w Centrum Informacji Naukowej i Bibliotece Akademickiej w Katowicach odbyło się seminarium naukowe „Przestrzenie humanistyki cyfrowej – korpusy diachroniczne polszczyzny". Wydarzenie zostało zorganizowane przez Instytut Języka Polskiego im. Ireny Bajerowej Uniwersytetu Śląskiego w Katowicach we współpracy z Instytutem Slawistyki na Uniwersytecie Jana Gutenberga w Moguncji oraz Fundacją Wiedzy i Dialogu Społecznego „Agere Aude". Bezpośrednim motywem, który skłonił organizatorów do zainicjowania przedsięwzięcia, było dostrzeżenie potrzeby gromadzenia i popularyzowania językowych danych diachronicznych w przestrzeni wirtualnej. Stworzenie platformy gromadzącej dostępne dane diachroniczne w sferze cyfrowej to ideał i zarazem wyzwanie, do którego realizacji pierwszym krokiem było przygotowanie seminarium. Zamysł jego zorganizowania zrodził się również z faktu dostrzeżenia nowych możliwości otwierających się przed językoznawcą wraz z postępującą digitalizacją językowych źródeł historycznych. Rola elektronicznych źródeł – w przekonaniu zarówno organizatorów, jak i zaproszonych gości – jest nieprzeceniona, ponieważ wpływa na jakość prowadzonych badań i znacznie skraca czas ich trwania.

Sesję referatową seminarium otworzyła w imieniu organizatorów prodziekan ds. promocji i współpracy z otoczeniem – Magdalena Pastuch. Pierwszym prelegentem był Włodzimierz Gruszczyński, który w swoim odczycie *Tagset barokowy – problemy opracowania zestawu kategorii morfologicznych i ich wartości na potrzeby Elektronicznego korpusu tekstów polskich XVII i XVIII w. (do 1772 r.)* podzielił się ze słuchaczami doświadczeniem tworzenia znaczników morfosyntaktycznych w projekcie „Elektronicznego korpusu tekstów polskich XVII i XVIII w. (do 1772 r.)". Jako drugi został wygłoszony referat poświęcony problemom ręcznej anotacji dużego korpusu tekstów. Sebastian Żurowski oraz Paulina Rosalska skoncentrowali się na omówieniu procesu pozyskiwania anotatorów i sposobu koordynowania zespołu oraz weryfikacji wyników pracy. Wskazali również na znaczącą rolę klarownej instrukcji, którą anotator musi otrzymać, rozpoczynając pracę z korpusem. Tematykę wywodu podtrzymała Imke Mendoza w referacie *Idealny korpus diachroniczny*. W swojej wypowiedzi wyróżniła kilkanaście istotnych (z różnych punktów widzenia) cech korpusu idealnego. Na początku prelegentka wskazała na występujące w przestrzeni cyfrowej problemy, które spotyka językoznawca, przeszukując korpusy, takie jak ograniczona liczba danych, duża wariantywność, znacząca rola kontekstu, heterogeniczność zabytków. Druga część odczytu poświęcona była cechom przypisywanym idealnemu korpusowi diachronicznemu. Pośród nich uwzględnione zostały między innymi: efektywne wyszukiwanie, nieskomplikowana obsługa, wyodrębnienie możliwych wariantów tekstu, ujednoznacznienie wielofunkcyjnych i polisemicznych elementów. Wyliczenia składowych idealnego korpusu zamknęły pierwszą część sesji referatowej. Po niej nastąpiła dyskusja, w której głosy dotyczyły przede

wszystkim wariantywności występującej w korpusach. Skonkludowano przy tym, iż lepiej, by ich użytkownik miał w rekordach nadmiar informacji aniżeli ich niedomiar. Dostrzeżono ponadto, że wiążąca dla twórców korpusu jest wizja gramatyki oraz sposób przypisywania znaczników morfosyntaktycznych; do tych informacji użytkownik powinien mieć umożliwiony dostęp. Na koniec podkreślono, że aktualnie nie występują na rynku pracy zawodowi anotatorzy tekstów, a ich rola, choć ważna, jest niedoceniana.

Po przerwie sesję rozpoczął Roland Meyer referatem *Polski korpus diachroniczny PolDi: Aktualny stan i perspektywy rozwoju*. Prelegent omówił sposób funkcjonowania korpusu PolDi oraz założenia przyjęte w trakcie jego powstawania. Zasygnalizował, że dalsze prace nad korpusem powinny zmierzać w kierunku disambiguacji pojęć i postępowej korekty danych.

Zagadnienie idealnego korpusu diachronicznego podjęła również kolejna referentka – Agnieszka Słoboda. Mówczyni zawęziła przedmiot swoich rozważań wyłącznie do tych korpusów diachronicznych, które dotyczą języka urzędowego, a ściślej – prawniczego. Podkreśliła także, że w założeniu w korpusie urzędniczym (np. *Rot sądowych*) należy zawsze uwzględniać kontekst wypowiedzi, badacz powinien mieć dostęp do podstawy tłumaczeniowej tekstów, a przede wszystkim wskazane jest, by teksty opierały się na źródle-oryginale, a nie na opracowaniach naukowych.

W kolejnym wystąpieniu *Problemy opisu dawnej fleksji na materiale szesnastowiecznych przekładów Ewangelii* Alina Kępińska i Izabela Winiarska-Górska przedstawiły na zgromadzonym materiale (staropolskich leksemach) różne sposoby odczytywania form fleksyjnych. W referacie prelegentki odniosły się do zakończonego w 2016 roku projektu poświęconego renesansowym tłumaczeniom Ewangelii. Badaczki zwróciły uwagę na portal www.ewangelie.uw.edu.pl, który gromadzi w wersji cyfrowej edycje krytyczne dziesięciu staropolskich translacji Ewangelii.

Ostatni referat – wygłoszony przez Tomasza Mikę – poświęcony był problemom badania staropolskiego języka religijnego z perspektywy twórcy i użytkownika korpusu oraz wskazywał na istotność uwzględniania w anotacji zachodzących procesów historycznych.

Po przerwie obiadowej miała miejsce praktyczna część seminarium. W pomieszczeniach CINiB-y odbywały się warsztaty, które zgromadziły językoznawców zainteresowanych zastosowaniem narzędzi cyfrowych w badaniach języka historycznego. W sali dydaktycznej Mariusz Leńczuk poprowadził warsztaty pt. *Piętnastowieczne przekłady Nowego Testamentu – elektroniczna konkordancja staropolska. Internetowa baza danych – Od pomysłu do realizacji. Źródła, wyzwania, perspektywy* i zaprezentował liczne możliwości, jakie niesie elektroniczna konkordancja staropolska. Zachęcił również zgromadzonych do wykorzystania potencjału tkwiącego w dostępnym narzędziu. W tym samym czasie odbywały się warsztaty poświęcone zastosowaniu równoległych korpusów synchronicznych w językoznawstwie diachronicznym. Ruprecht von Waldenfels przedstawił uczestnikom korpus ParaViz i możliwości wykorzystania zgromadzonych w nim danych. Następnie w sesji warsztatowej miała miejsce prezentacja „Elektronicznego korpusu tekstów polskich XVII i XVIII w. (do 1772 r.)", podczas której prowadząca Renata Bronikowska objaśniła znakowanie morfosyntaktyczne zastosowane w korpusie, a także wskazała na zróżnicowane możliwości wyszukiwarki.

W drugim dniu seminarium wygłoszono pięć referatów. Pierwszy poświęcony był analizatorowi fleksyjnemu polszczyzny lat 1830–1918. Z możliwościami omawianego narzędzia internetowego zaznajomiła słuchaczy Magdalena Derwojedowa. Jako kolejny wystąpił Piotr

Sobotka, który w wygłoszonym odczycie wskazał sposoby anotacji staropolskich wyrażeń w korpusie języka dawnego. Następnie wspomniany już von Waldenfels przedstawił problem ekspansji przyimka *do* w językach północnosłowiańskich na podstawie równoległych korpusów diachronicznych. Z kolei Magdalena Król w wystąpieniu (przygotowanym wraz z Maciejem Ederem) zaznajomiła słuchaczy ze specyfiką funkcjonowania imiesłowu przysłówkowego w XVI i XVII wieku. Badania z użyciem korpusu diachronicznego pozwalają odnotować – jak zaznaczała prelegentka – spadek produktywności omawianego imiesłowu. Ponadto mówczyni wyjaśniła słuchaczom, w jaki sposób, wykorzystując tzw. metodę TTR (Type-Token-Ratio), ustalić relację pomiędzy wyrazami podzielnymi słowotwórczo a tokenami.

Po odczytach miała miejsce burzliwa dyskusja dotycząca przede wszystkim terminu *jednostka historyczna*, której wprowadzenie postulował Sobotka. Zastanawiano się ponadto nad rolą metadanych w korpusie, zagadnieniem zanikania imiesłowów w polszczyźnie oraz koniecznością lematyzacji powstających korpusów. Żywa wymiana myśli w wielu przypadkach rzuciła nowe światło na sporne kwestie i poszerzyła perspektywę ich oglądu.

Po przerwie głos zabrały główne organizatorki seminarium: Magdalena Pastuch, Beata Duda, Karolina Lisczyk, Katarzyna Sujkowska-Sobisz. Ich referat *Diachroniczne korpusy polszczyzny – o potrzebie tworzenia konstelacji językowych hurtowni danych*, dotyczył potrzeby budowania platformy gromadzącej informacje o dostępnych korpusach diachronicznych, ponieważ – jak zapewniały prelegentki – takiego miejsca w przestrzeni wirtualnej jeszcze nie ma. Co istotne, postulowane narzędzie miałoby przyjmować perspektywę użytkownika, a nie twórcy korpusu. Platforma stanowiłaby dla językoznawcy swoistą bazę ułatwiającą efektywne prowadzenie badań. Następnie zaprezentowano wszystkie aktualnie dostępne w internecie diachroniczne bazy danych o różnym charakterze. Przedstawiono nie tylko projekty ukończone, ale też te, które są w fazie realizacji. Kolejnym punktem seminarium był panel dyskusyjny. Wymiana poglądów prowadzona przez Bjoerna Wiemera została zdominowana przez zagadnienie różnorodnych sposobów funkcjonowania elektronicznych źródeł językowych poświęconych językowi polskiemu i jego historii. Idea powstania językowej hurtowni danych zyskała aprobatę, a przyszłe kroki w realizacji pomysłu powinny dotyczyć już kwestii związanych z praktyczną stroną przedsięwzięcia.

Seminarium zakończyły warsztaty naukowe, które dotyczyły wykorzystania w badaniach infrastruktury naukowej CLARIN. Miały one charakter otwarty (zaadresowane były do studentów, doktorantów i pracowników naukowych), a wstęp na nie był wolny. Prowadzący – Maciej Piasecki i Mariusz Oleksy – w pierwszej części zapoznali uczestników z infrastrukturą naukową technologii językowych CLARIN. W drugiej części zaprezentowano możliwości przetwarzania danych w repozytorium DSpace. W szczególności skupiono się na metadanych w procesie publikacji, licencjach, formatach plików źródłowych oraz na automatycznym przetwarzaniu wyników. Tematem drugiego bloku warsztatów był webowy system do konstrukcji korpusów tekstowych. Szczegółowo poruszono takie zagadnienia, jak import korpusu, struktura i zarządzanie korpusem, opisywanie dokumentów, anotacja, statystyki korpusu, listy frekwencyjne słów i anotacji czy możliwości eksportu anotowanego korpusu. Część ćwiczeniowa stanowiła ostatni punkt warsztatów i jednocześnie zamykała seminarium poświęcone korpusom diachronicznym.