RECENZJE

POLONICA XXXV
PL ISSN 0137-9712

Praktyczny przewodnik po korpusach języków słowiańskich, red. Milena Hebal-Jezierska, Wydział Polonistyki Uniwersytetu Warszawskiego, Warszawa 2014, s. 231.

Praktyczny przewodnik po korpusach języków słowiańskich to praca zbiorowa opublikowana nakładem Wydziału Polonistyki Uniwersytetu Warszawskiego w 2014 roku. Poza wydaniem w formie papierowej dostępna jest również w Internecie bezpłatna wersja w formie pliku PDF[1]. Prezentacja publikacji odbyła się w czerwcu 2014 r. w Warszawie na konferencji „Grammar and Corpora" zorganizowanej przez Instytut Slawistyki Zachodniej i Południowej Uniwersytetu Warszawskiego oraz Fundację Slawistyczną.

Książka kierowana jest do osób, które interesują się zagadnieniami językoznawstwa korpusowego lub zajmują się dydaktyką języków słowiańskich. We Wstępie „Przewodnika" wspomniano, że może być on przydatny dla „językoznawców, tłumaczy, lektorów oraz uczestników kursów studiujących języki słowiańskie".

Wprowadzenie do rozdziałów szczegółowych, poświęconych konkretnym językom słowiańskim, stanowi Rozdział 1, uporządkowano i wyjaśniono w nim terminy (wraz z synonimami), których używa się w charakterystykach poszczególnych korpusów. Rozdział ma charakter i układ słownika. W alfabetycznej kolejności opisane zostały najważniejsze pojęcia z zakresu lingwistyki korpusowej. Uporządkowany opis znaczeń i zakres użycia terminów pojawiających się w książce ułatwia jej lekturę.

Kolejne rozdziały stanowią cenne źródło informacji oraz ocenę korpusów języków słowiańskich — zarówno tych stworzonych jakiś czas temu, jak i tych, które są w trakcie powstawania (np. korpus języka macedońskiego).

Każdy z rozdziałów stanowi omówienie elektronicznego zbioru tekstów jednego z języków słowiańskich. W opracowaniu uwzględniono charakterystyki korpusów następujących języków: polskiego, czeskiego, słowackiego, dolnołużyckiego, górnołużyckiego, chorwackiego, serbskiego, słoweńskiego, bułgarskiego, macedońskiego, rosyjskiego, ukraińskiego i białoruskiego. Dwa ostatnie rozdziały przedstawiają korpusy równoległe.

Opisy w „Przewodniku" po korpusach mają swoją uporządkowaną, ustaloną przez zespół autorów strukturę. Autorami rozdziałów są osoby, dla których elektroniczne zbiory tekstów są narzędziami badawczymi. Są wśród nich także współtwórcy korpusów.

[1] Adres strony: http://www.iszip.uw.edu.pl/files/pdf/praktyczny_przewodnik.pdf.

W książce przeważają teksty napisane oryginalnie po polsku (przez Polaków lub badaczy działających w Polsce), ale są także trzy rozdziały tłumaczone z języków obcych.

Wstępne informacje w każdym rozdziale obejmują przedstawienie korpusu lub korpusów danego języka, wskazanie nazw ośrodków odpowiedzialnych za powstanie i kształt korpusów, określenie momentu ich opracowania, prezentację krótkiej historii i charakterystyki oraz uwzględnienie innych danych odnoszących się do podstawowych założeń organizacji projektu. Podaje się także informację o liczebności i rodzajach korpusów i podkorpusów. Następnie wymienia się adresy internetowe i inne wskazówki ułatwiające do nich dostęp. W niektórych rozdziałach zamieszczono również charakterystykę danego języka słowiańskiego, piśmiennictwa lub jego struktury, np. w rozdziale poświęconym korpusowi języka serbskiego.

Cześć pierwsza każdego rozdziału to opis prezentujący strukturę korpusu. Uwzględnia się w niej informację o sposobach pozyskiwania i segmentacji tekstów, rodzajach tekstów zaimplementowanych w korpusie, ich charakterze genologicznym, poziomie zrównoważenia stylowego, gatunkowego, tematycznego i chronologicznego. Kolejne informacje dotyczą anotacji zewnętrznej (odnosi się ona do metadanych tekstu, takich jak: autor, tytuł, data powstania/publikacji itp.) oraz anotacji wewnętrznej (tzn. lematyzacji i tagowania, czyli sposobu nadawania segmentom tekstu informacji gramatycznych, w niektórych korpusach także semantycznych i dodatkowych, np. słowotwórczych). W poszczególnych rozdziałach opisuje się narzędzia wbudowane w korpusy, programy służące do automatycznego znakowania tekstów, a także wskazuje się na naukowe opracowania gramatyczne, na których oparty jest system znaczników. W rozdziałach opisujących anotację korpusów zwraca się uwagę na powtarzające się problemy wynikające ze znakowania tekstów (najczęściej wskazuje się stopień dezambiguacji, problemy homonimii, ortografii oraz rozwiązania znakowania interpunkcji).

Następnym elementem opisu uwzględnionym w każdym z rozdziałów jest zbiór zaleceń dotyczących sposobu korzystania z korpusu. Niemal w każdym z opisów podaje się podstawowe informacje o sposobie wyszukiwania jednostek i przykładowe zapytania, do których dołączone są zrzuty ekranu z wynikami wyszukiwania. Część rozdziału poświęcona metodom korzystania z korpusów może stanowić praktyczny zestaw pomocy i wskazówek dla osób zainteresowanych pracą na nieznanym wcześniej korpusie. Autorzy często podpowiadają**,** w jaki sposób uzyskać wiarygodne wyniki oraz jakich unikać błędów, by nie doprowadzić do fałszywych wniosków bądź zbytnich uogólnień wynikających ze wstępnych analiz wyników wyszukiwania i automatycznych statystyk.

W rozdziałach poświęconych większym projektom korpusowym uwzględnia się także charakterystykę poszczególnych podkorpusów, wyszukiwarek i innych funkcji oraz narzędzi mających zastosowanie we współczesnych badaniach lingwistycznych. W wypadku korpusów bardziej rozbudowanych strukturalnie (takich jak korpus języka czeskiego lub rosyjskiego) osobną charakterystykę sporządzono dla korpusu głównego, zaś osobne opisy stanowią korpusy specjalistyczne i równoległe, funkcjonujące w obrębie tego samego projektu.

W zakończeniu każdego z rozdziałów autorzy podają informację o zastosowaniu korpusu w badaniach nad językiem. Wymienia się publikacje zawierające wyniki badań nad gramatyką konkretnego języka lub wskazuje słowniki, dla których korpus tekstów stanowi bazę materiałową. Wielu autorów wymienia wady i zalety korpusu**,** proponując dziedziny i zagadnienia naukowe, w których opisywany korpus może mieć zastosowanie.

Dwa ostatnie rozdziały, w których omawiane są korpusy równoległe, różnią się strukturą opisu od rozdziałów podejmujących zagadnienie korpusów narodowych. Jeden z rozdziałów został poświęcony korpusowi ParaSol i Korpusowi polsko-rosyjskiemu UW, natomiast drugi poświęcono korpusowi InterCorp.

We wstępie pierwszego z rozdziałów jego autor, Marek Łaziński, przywołał krótką historię równoległego publikowania tekstów oraz podał współczesne przykłady korpusów wielojęzycznych. W opisach korpusów języków narodowych podkreśla się ich rolę w tworzeniu gramatyk, natomiast w wypadku korpusów wielojęzycznych — zwraca się uwagę na wartość komputerowych narzędzi w pracach tłumaczeniowych. W opisach wszystkich trzech korpusów równoległych wymieniono języki tekstów włączonych do korpusu. W charakterystyce korpusu ParaSol i InterCorp dodatkowo wskazano możliwości wyboru języka, w którym zadaje się zapytanie, oraz kolejnych języków obcych. Podobnie jak w rozdziałach o poszczególnych korpusach języków słowiańskich, opisuje się strukturę zbioru, anotację, metody prowadzenia analiz, przykładowe zapytania oraz ocenę funkcjonalności interfejsów. Rozdziały kończą uwagi odnoszące się do zalet i wad omawianych korpusów.

Autorzy wszystkich rozdziałów niejednokrotnie w sposób krytyczny oceniają wartość analizowanego zbioru tekstów. Dzielą się z czytelnikiem uwagami dotyczącymi ich budowy, zasobu oraz praktycznych zastosowań. Oceny te budują na podstawie własnych doświadczeń zdobytych w czasie przeprowadzonych przez siebie analiz korpusowych.

Praktyczny przewodnik po korpusach języków słowiańskich jest pracą niezwykle wartościową, ponieważ nie tylko zbiera wiele charakterystyk poszczególnych korpusów, ale jest również cenny ze względu na swą użyteczność. Można powiedzieć, że dzięki swojej uporządkowanej formie i włączeniu w opis wskazówek, a nierzadko instrukcji wyszukiwania i korzystania z podkorpusów i dodatkowych, mniej znanych ich funkcji, jest publikacją, która z pewnością wielu zainteresowanym pozwoli lepiej się rozeznać w dziedzinie lingwistyki komputerowej i korpusowej w obrębie języków słowiańskich. Jest istotnym dokumentem, który przedstawia obecny stopień zaawansowania badań w zakresie językoznawstwa korpusowego w różnych krajach. Opisy rozbudowanych i zaawansowanych technologicznie korpusów mogą stać się źródłem inspiracji dla twórców korpusów dopiero powstających, zaś uwagi krytyczne, wynikające ze spostrzeżeń badaczy, mogą inspirować do opracowywania doskonalszych i bardziej funkcjonalnych korpusów tekstów, w których uwzględnia się potrzeby większej grupy użytkowników.

Monika Kasza
Instytut Języka Polskiego PAN, Kraków