Korpusomat — narzędzie do tworzenia przeszukiwalnych korpusów języka polskiego

Witold Kieraś, Łukasz Kobyliński, Maciej Ogrodniczuk
Instytut Podstaw Informatyki PAN

Streszczenie:

Korpusomat to internetowe narzędzie ułatwiające użytkownikowi samodzielne tworzenie i wykorzystywanie korpusów do badań językoznawczych. Narzędzie umożliwia przesłanie zestawu plików tekstowych wraz z metadanymi, a następnie zlecenie ich automatycznej analizy lingwistycznej. Powstały korpus można następnie pobrać i badać na własnym komputerze, używając wyszukiwarki korpusowej Poliqarp. Zaletą aplikacji jest brak konieczności znajomości szczegółów technicznych rozwiązań informatycznych wykorzystywanych do analizy tekstów – wszystkie komponenty analityczne zainstalowane są na serwerze, który kompiluje korpus, a zadania użytkownika ograniczają się do przygotowania tekstów i wprowadzania zapytań przeszukujących korpus. Analiza lingwistyczna tworzonego korpusu polega na wykryciu wszystkich interpretacji fleksyjnych poszczególnych słów wraz z ich formami hasłowymi oraz wyborze interpretacji najbardziej prawdopodobnej w określonym kontekście. Na tej podstawie wyszukiwarka korpusowa może wykonywać zapytania dotyczące segmentacji tekstu, form podstawowych, znaczników fleksyjnych, wieloznaczności i dezambiguacji, a także ograniczać zapytania za pomocą metadanych. Wynikiem wyszukiwania lingwistycznego jest konkordancja, czyli zestaw wszystkich wystąpień danego wyrazu w korpusie wraz z jego najbliższym kontekstem. Poliqarp pozwala także na zliczanie frekwencji określonych wyrazów oraz stosowanie podstawowych miar statystycznych niezbędnych w badaniach kwantytatywnych.
Korpusomat łączy istniejące aplikacje wykorzystywane w pracach nad Narodowym Korpusem Języka Polskiego: analizator morfologiczny Morfeusz, tager ujednoznaczniający Concraft oraz wyszukiwarkę korpusową Poliqarp, której język zapytań znany jest użytkownikom NKJP i dzięki temu pozwala np. na łatwe porównywanie wyników uzyskiwanych w dużym korpusie ogólnym (np. zrównoważonym podkorpusie NKJP) z wynikami z małych, samodzielnie zgromadzonych korpusów specjalistycznych poświęconych konkretnemu autorowi, tematyce czy konkretnemu rodzajowi prasy, które można stworzyć dzięki prezentowanej aplikacji. Niewątpliwą zaletą Korpusomatu jest obsługiwanie popularnego formatu plików ePUB, w którym dystrybuowana jest zdecydowana większość dostępnych na rynku ebooków i prasy elektronicznej.