

CLARIN-PL

Technologie IT w naukach społecznych i humanistycznych

Politechnika Wrocławska jest liderem konsorcjum polskiej części ogólnoeuropejskiej struktury naukowej CLARIN, obejmującej cyfrowe archiwa tekstowe i narzędzia do ich automatycznej analizy. Umożliwia to badaczom nauk humanistycznych i ścisłych pracę z dużymi zbiorami tekstów. Stanowi również olbrzymią pomoc dla naukowców z innych obszarów nauki. Z nowoczesnych narzędzi opracowanych w ramach projektu mogą korzystać także podmioty komercyjne. CLARIN-PL jest realizowany dzięki dofinansowaniu ze środków Programu Operacyjnego Inteligentny Rozwój, które otrzymał w ramach konkursu ogłoszonego przez Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy.

■ CLARIN-PL z certyfikatem Core Trust Seal

CLARIN to złożony system centrów technologicznych dający dostęp do zasobów, narzędzi i aplikacji badawczych. Jego głównym celem jest umożliwienie badań naukowych w obszarze nauk humanistycznych i społecznych z wykorzystaniem technologii językowej. Dzięki otwartemu dostępowi do większości komponentów Centrum Technologii Językowych, wpływa on również na inne obszary nauki. Z nowoczesnych narzędzi opracowanych w ramach projektu mogą korzystać także podmioty komercyjne, które potrzebują aplikacji do przetwarzania języka naturalnego. Warto podkreślić, że Centrum Technologii Językowych posiada światowy certyfikat repozytorium danych badawczych Core Trust Seal.

– Głównym celem CLARIN-PL jest dostarczenie polskim naukowcom wszelkich środków potrzebnych do prowadzenia badań wymagających wykorzystania metod przetwarzania dużych ilości tekstów. Nasza infrastruktura jest unikatowa na skalę światową, ponieważ powstała specjalnie z myślą o języku polskim i dostarcza autorskich rozwiązań zapewniających precyzyjne wyniki dla badań na tekstach sporządzonych w tym języku. Nasze działania mają na celu udostępnianie

zasobów językowych oraz elektronicznych narzędzi do automatycznego przetwarzania języka naturalnego badaczom z różnych dyscyplin naukowych, a w szczególności z dziedziny nauk humanistycznych i społecznych – mówi dr hab. inż. Maciej Piasecki z Politechniki Wrocławskiej, kierownik projektu CLARIN-PL.

■ Nowoczesne e-Humanities

CLARIN-PL wychodzi naprzeciw wyzwaniom i perspektywom badawczym w dziedzinie e-Nauki, zwłaszcza tzw. e-Humanities. E-Nauka jako nowy nurt metodologiczny zakłada szerokie zastosowanie technologii wspierających użytkownika w posługiwaniu się dużymi zbiorami tekstów. W Centrum Technologii Językowych powstają programy, które pomocne będą m.in. w ustalaniu autorstwa tekstów anonimowych, określaniu profilu psychologicznego autora, wydobyciu z tekstów wiedzy i informacji, badaniu powiązań w biznesie, polityce i nauce, anonimizacji i normalizacji tekstu oraz wielu innych.

■ Słowniec - drugi co do wielkości wordnet na świecie

Centrum Technologii Językowych udostępnia także zasoby językowe, np. Korpus Politechniki Wrocławskiej, Korpus Dyskursu Parlamentarnego oraz Słowniec – drugi co do wielkości wordnet na świecie. Służą one do uczenia programów „rozumienia” języka naturalnego. Z zasobów tych mogą korzystać również indywidualni użytkownicy. Tylko w pierwszym kwartale 2021 r. narzędzia IT, opracowane w ramach CLARIN-PL, znalazły ponad 120 nowych użytkowników: indywidualnych, instytucjonalnych i komercyjnych (zarówno polskich, jak i zagranicznych). Przykładowo, dane ze Słownioci zostały użyte przez Open Multilingual Wordnet i połączone z danymi z ponad 60 języków świata. Natomiast WordnetLoom – system do edycji leksykonów semantycznych, był używany przez zespoły z RPA, Portugalii i Danii. Został on

także wykorzystany w opracowaniu wielojęzycznego słownika jkdysz. Aplikacje Centrum Technologii Językowych wykorzystuje również Centrum Analiz Strategicznych przy Kancelarii Prezesa Rady Ministrów RP, które używa tych narzędzi do analizy sentymentu i normalizacji tekstu w monitorowaniu dyskusji publicznej.

Projekt CLARIN-PL realizuje konsorcjum: Instytut Podstaw Informatyki PAN, Uniwersytet Łódzki, Instytut Sławiastyki PAN, Uniwersytet Wrocławski oraz Politechnika Wrocławska (lider projektu), a także 27 podmiotów komercyjnych, które zasiłowały projekt wkładem własnym. Projekt jest współfinansowany ze środków POIR. Instytucją Wdrażającą Działanie 4.2. „Rozwój nowoczesnej infrastruktury badawczej sektora nauki” w ramach IV Osi Priorytetowej POIR „Zwiększenie potencjału naukowo-badawczego” jest Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy.



„Środki, które przekazaliśmy w ramach ogłoszonego przez OPI konkursu, zostały bardzo dobrze wykorzystane. CLARIN-PL oferuje kilkadziesiąt webserwisów, obsługuje miliony zapytań i wspiera setki użytkowników rocznie. Skala wykorzystania Centrum Technologii Językowych wzrasta każdego roku. Większość opracowanych narzędzi ma otwarte licencje i przyciągnęła do tej pory ponad sto firm.”

Joanna Kuszlik-Cichosz
zastępca dyrektora ds. administracji w Ośrodku Przetwarzania Informacji – Państwowym Instytucie Badawczym