

Projekt „CLARIN – Wspólne zasoby językowe i infrastruktura technologiczna” (Nr FENG.02.04-IP.04-0004/24) finansowany z Drugiego Priorytetu Programu Funduszy Europejskich dla Nowoczesnej Gospodarki 2021–2027 (FENG)

Instytut Sławistyki PAN realizuje w ramach konsorcjum CLARIN-PL-BIZ projekt „CLARIN – Wspólne zasoby językowe i infrastruktura technologiczna” (Nr FENG.02.04-IP.04-0004/24) finansowany z Drugiego Priorytetu Programu Funduszy Europejskich dla Nowoczesnej Gospodarki 2021–2027 (FENG). Całkowita wartość projektu wynosi 61.141.241,03 zł, w tym wysokość wkładu Funduszy Europejskich to 48.669.499,44 zł. Do prac nad projektem Instytut Sławistyki przystąpił 1 stycznia 2025 r.

W skład konsorcjum naukowego **CLARIN-PL-BIZ** wchodzi następujące instytucje badawcze:

- **Politechnika Wrocławska** (Lider Projektu; osoba odpowiedzialna: prof. PWr Maciej Piasecki)
- **Instytut Podstaw Informatyki PAN** (prof. IPI PAN Maciej Ogrodniczuk)
- **Instytut Sławistyki PAN** (dr hab. Roman Roszko, prof. IS PAN)
- **Uniwersytet Łódzki** (prof. UŁ Piotr Pęzik)
- **Uniwersytet Wrocławski** (prof. Adam Pawłowski)

Cele CLARIN-PL-BIZ

Celem Projektu jest rozwój narzędzi i rozwiązań wspierających zarówno naukowców, jak i przedsiębiorców w dostępie do zaawansowanych zasobów językowych i technologicznych. Trzonem wszystkich zadań projektowych jest rozszerzenie infrastruktury badawczej CLARIN-PL powstałej w wyniku poprzednich projektów: CLARIN-PL-BIZ (POIR 4.2, zakończony w roku 2023), CLARIN-PL I, CLARIN-PL II, CLARIN-PL III, CLARIN-PL IV oraz obecnie realizowanego CLARIN-PL V (2024–2026).

Prace projektowe są zogniskowane na dalszym usprawnianiu wsparcia dla badaczy w obszarach badań naukowych i działań innowacyjnych. Obejmują między innymi:

- **Analitykę zbiorów** – rozwój narzędzi i metod do analizy dużych zbiorów danych.
- **Dane językowe** – rozbudowę i ulepszenie istniejących danych językowych oraz zwiększenie dostępności do zawartej w nich wiedzy.
- **Wydobywanie informacji** – rozwinięcie technik wydobywania kluczowych informacji.
- **Rozwiązania spersonalizowane** – tworzenie rozwiązań uwzględniających różnorodne perspektywy postrzegania.
- **Naturalna i efektywna komunikacja** – opracowanie rozwiązań dla naturalnej i skutecznej komunikacji w języku naturalnym.
- **Tworzenie korpusów językowych** – tworzenie baz danych z tekstami w różnych językach.

Zadania Zespołu Instytutu Sławistyki PAN w CLARIN-PL-BIZ

Zespół Semantyki i Lingwistyki Komputerowej Instytutu Sławistyki PAN (Zespół SiLK) pod kierunkiem dr. hab. Romana Roszko, prof. IS PAN, w ramach Projektu realizuje szereg zadań, stosując nowoczesne technologie do głębszego zrozumienia języka oraz poprawy jakości komunikacji w przestrzeni cyfrowej. Skupia się na tworzeniu zaawansowanych danych ewaluacyjnych i testowych (referencyjnych), przeznaczonych do oceny i testowania generatywnych modeli językowych z dziedzin takich jak: weterynaria, indoeuropeistyka, literatura, historia, dialektologia polska i inne.

Praktyczne zastosowanie tworzonych przez Zespół SiLK organicznych zasobów danych tekstowych to przede wszystkim wykorzystanie ich w benchmarkach do głębokiego testowania jakości generatywnych

modeli językowych. Te dane są kluczowe dla ewaluacji i rozwoju nowoczesnych rozwiązań w dziedzinie przetwarzania języka naturalnego.

Skład Zespołu Semantyki i Lingwistyki Komputerowej Instytutu Slawistyki PAN:

- Roman Roszko (kierownik, lingwista),
- Karolina Bernaś (lekarka weterynarii),
- Tomasz Bernaś (informatyk i lingwista),
- Magdalena Gardias (lingwistka),
- Karol Kościelniak (historyk),
- Piotr Szatkowski (lingwista),
- Valéry Trân Thiên (informatyk i lingwista).