CLARIN ERIC (Common Language Resources & Technology Infrastructure, pol. Wspólne zasoby językowe i infrastruktura technologiczna) to ogólnoeuropejska infrastruktura naukowa, udostępniająca zasoby i narzędzia językowe dla wszystkich języków europejskich w ramach wspólnej infrastruktury badawczej, stanowiącej narzędzie pracy naukowców reprezentujących nauki humanistyczne i społeczne (NHiS).

CLARIN-PL to polskie konsorcjum złożone z sześciu jednostek naukowych, wnoszące wkład w rozwój i utrzymanie CLARIN ERIC, działające szczególnie w zakresie technologii dla języków polskiego i wybranych języków słowiańskich, w powiązaniu z innymi językami o zasięgu światowym. Polska jest jednym z siedmiu członków-założycieli.
CLARIN ERIC zbiera w sieciowym systemie rozproszone zasoby i narzędzia językowe, usługi sieciowe i aplikacje badawcze do pracy ze zbiorami tekstów lub nagraniami i udostępnia je wg wspólnych standardów opisu i dostępu (jednolity system logowania federacyjnego, otwarty dostęp i licencje, standard metadanych CMDI). Podstawowe funkcje przeszukiwania zasobów są zapewnione na poziomie centralnym, a usługi i aplikacje są wytwarzane i oferowane przez narodowe konsorcja w ramach połączonej infrastruktury. CLARIN rozwija centralną platformę dostępu do narzędzi językowych i aplikacji badawczych oraz pełni kluczową rolę w budowie EOSC w obszarze klastra technologicznego NHiS o nazwie SSH Open Marketplace.

Celem projektu jest wypełnienie zobowiązań Polski wynikających z członkostwa w CLARIN ERIC poprzez wymienione poniżej działania:

1) Utrzymanie polskiej części infrastruktury CLARIN ERIC - CLARIN-PL - zbudowanej w latach 2013-18, czyli utrzymanie w nieustającej gotowości do użycia i na wysokim poziomie dostępności i użyteczności Centrum Technologii Językowych CLARIN-PL (CTJ) jako jedynego polskiego centrum CLARIN typu B (centrum technologiczne CLARIN ERIC), zapewniającego wybrane funkcje jako centrum typu A (tj. świadczącego usługi na rzecz centralnej funkcjonalności CLARIN ERIC).

2) Utrzymanie odpowiedniego wysokiego poziomu integracji infrastruktury CLARIN-PL z ewoluującą rozproszoną infrastrukturą CLARIN ERIC oraz EOSC, dziedzinowymi klastrami technologicznymi w EOSC i innymi europejskimi infrastrukturami badawczymi.

3) Utrzymanie efektywnego poziomu wsparcia dla użytkowników CLARIN-PL, szczególnie utrzymanie polskiego Centrum Wiedzy CLARIN Technologii Językowej dla Języka Polskiego PolLinguaTec - certyfikowanego centrum CLARIN K, części Infrastruktury Wiedzy CLARIN.

4) Szerokie propagowanie wiedzy o CLARIN, w tym CLARIN-PL i wdrażanie CLARIN jako narzędzia badawczego, wśród naukowców i studentów, szczególnie w obszarze NHiS.

CLARIN-PL jest skupiony przede wszystkim na języku polskim - ważnym składniku polskiego dziedzictwa kulturowego, pozostającym jednak ciągle poza głównym obszarem zainteresowań badań międzynarodowych. Wysoki poziom integracji z infrastrukturą europejską oraz powiązanie z j. angielskim zapewniają bardzo dobrą widoczność i dostępność polskich rozwiązań na poziomie międzynarodowym, przeciwdziałając cyfrowemu wykluczeniu językowemu. Otwarte rozwiązania CLARIN-PL mają zastosowania w NHiS oraz innych dziedzinach, np. informatyce, medycynie, architekturze, przemyśle, administracji, bibliotekach, NGO. CLARIN tworzy narzędzia do przetwarzania zasobów dawnych,

dostępnych dotąd jedynie nielicznym specjalistom. Potwierdzeniem wagi rozwiązań w zakresie PJN jest uznanie technologii jęz. za integralny element promocji polszczyzny przez Radę Języka Polskiego, a przykładami stosowania współpraca z BN, AZON, UOKiK, KPRM, czy też KJ PAN.

Prace w ramach projektu realizowane są przez Zespół Semantyki i Lingwistyki Korpusowej Instytutu Slawistyki PAN (kierownik dr hab. Roman Roszko, prof. IS PAN). Zespół SiLK (dr hab. Roman Roszko, prof. IS PAN, mgr Tomasz Bernaś, prof. Andrius Utka, mgr Daniel Dziułka, dr Jewgienija Żejmo) prowadził prace nad pozyskaniem nowych zasobów wielojęzycznymi głównie z języków słowiańskich (polskiego, białoruskiego, ukraińskiego, czeskiego, słowackiego, bułgarskiego) i bałtyckich (litewskiego, łotewskiego) oraz germańskich (niemiecki) w dowolnej konfiguracji. Powstały pokaźne zasoby dwujęzyczne: litewsko-bułgarskie (2,84 GB w plikach TMX), litewsko-łotewskie (2,94 GB), polsko-bułgarskie (2,90 GB), polsko-czeskie (2,99 GB), polsko-litewskie (3,01 GB), polsko-łotewskie (2,99 GB), polsko-słowackie (3,05 GB) i inne.

Ze względu na obcojęzyczne wtrącenia (wewnątrzzdaniowe i międzysegmentalne) opracowano algorytm filtracji niepożądanych treści.