CLARIN
Common Language Resources and Technology Infrastructure

CLARIN-PL

CLARIN – rozproszony system technologii językowych dla różnych języków europejskich

Maciej Piasecki
CLARIN-PL
Politechnika Wrocławska
Instytut Informatyki
G4.19 Research Group
maciej.piasecki@pwr.wroc.pl
2013-04-17

Projekt CLARIN

- CLARIN =
  - Common Language Resources and Technology Infrastructure
  - *Wspólne zasoby językowe i infrastruktura technologiczna*
- Część
  - europejskiej mapy drogowej infrastruktury naukowej (European Roadmap for Research Infrastructures) ESFRI (European Strategy Forum on Research Infrastructures)
  - Polskiej Mapy Drogowej Infrastruktury Badawczej
- Cel
  - *zasobów i narzędzi językowych* dla wszystkich języków europejskich w ramach jednej wspólnej sieciowej infrastruktury naukowej
- Obszar działania: nauki humanistyczne i społeczne

Projekt CLARIN

- CLARIN ERIC
  - konsorcjum naukowe typu ERIC (*European Research Infrastructure Consortium*)
  - członkowie
    - Austria
    - Bułgaria
    - Czechy
    - Dania
    - Estonia
    - Holandia
    - Niemcy
    - *Dutch Language Union* (organizacja międzypaństwowa)
  - obserwatorzy
    - Norwegia

Podstawowe pojęcia

- Zasoby językowe
  - opisy języka naturalnego, które są sformalizowane w różnym stopniu
  - zbiory danych i bazy danych opisujące język naturalny oraz jego użycie
- Narzędzia językowe
  - programy komputerowe do przetwarzania tekstu i mowy na różnych poziomach analizy języka naturalnego
  - automatyczna analiza struktur językowych, np. analiza składniowa
  - zastosowania użytkowe, np. rozpoznawanie i klasyfikacja nazw własnych
- Technologia językowa = zasoby + narzędzia + infrastruktura
- Infrastruktura językowa
  - wspólna baza technologiczna zapewniająca połączenie zróżnicowanych narzędzi i zasobów językowych

Zasoby językowe

- Korpusy (duże zbiory) dokumentów tekstowych i nagrań mowy:
  - przykłady użycia (fragmenty, wypowiedzi lub całe dokumenty)
  - anotowane - opisane pod względem lingwistycznym w sformalizowany sposób (np. pod względem gramatycznym, czy też znaczenia)
- Słowniki
  - morfologiczne,
  - własności gramatycznych słów,
  - nazw własnych,
  - leksykony semantyczne, leksykalne sieci semantyczne,
  - wielojęzyczne słowniki itd.
- Gramatyki
- Inne zasoby
  - np. schematy anotacji oraz metadanych, funkcje podobieństwa semantycznego słów, listy częstościowe, modele językowe itd.

Narzędzia językowe

- Analizatory morfologiczne — rozpoznające znane słowa i przypisujące im opis własności gramatycznych
- Programy do ujednoznaczniania znaczeń słów w tekście
- Parsery
  - dokonujące analizy składniowej
  - oraz semantycznej tekstu
- Programy do rozpoznawania mowy i pisma ręcznego
- Programy do analizy znaczenia i struktury znaczeniowej tekstu
  - rozpoznawanie i klasyfikacja nazw własnych
  - rozpoznawanie powiązań anaforycznych
  - rozpoznawanie sytuacji
- itd.

Narzędzia językowe - przykład

Autorzy filmu stawiają tezę, że pierwszym człowiekiem, który postawił nogę na wierzchołku Ziemi był George Mallory.

Autor film stawiać teza , że pierwszy człowiek , który postawić noga na wierzchołek Ziemia być George Mallory .

Autor[subst.nom.pl.m1] film[subst.gen.sg.m3] stawiać[fin.pl.m1] teza[subst.gen.sg.f] , że[conj] pierwszy[adj.dat.sg.m1] człowiek[subst.dat.sg.m1] …

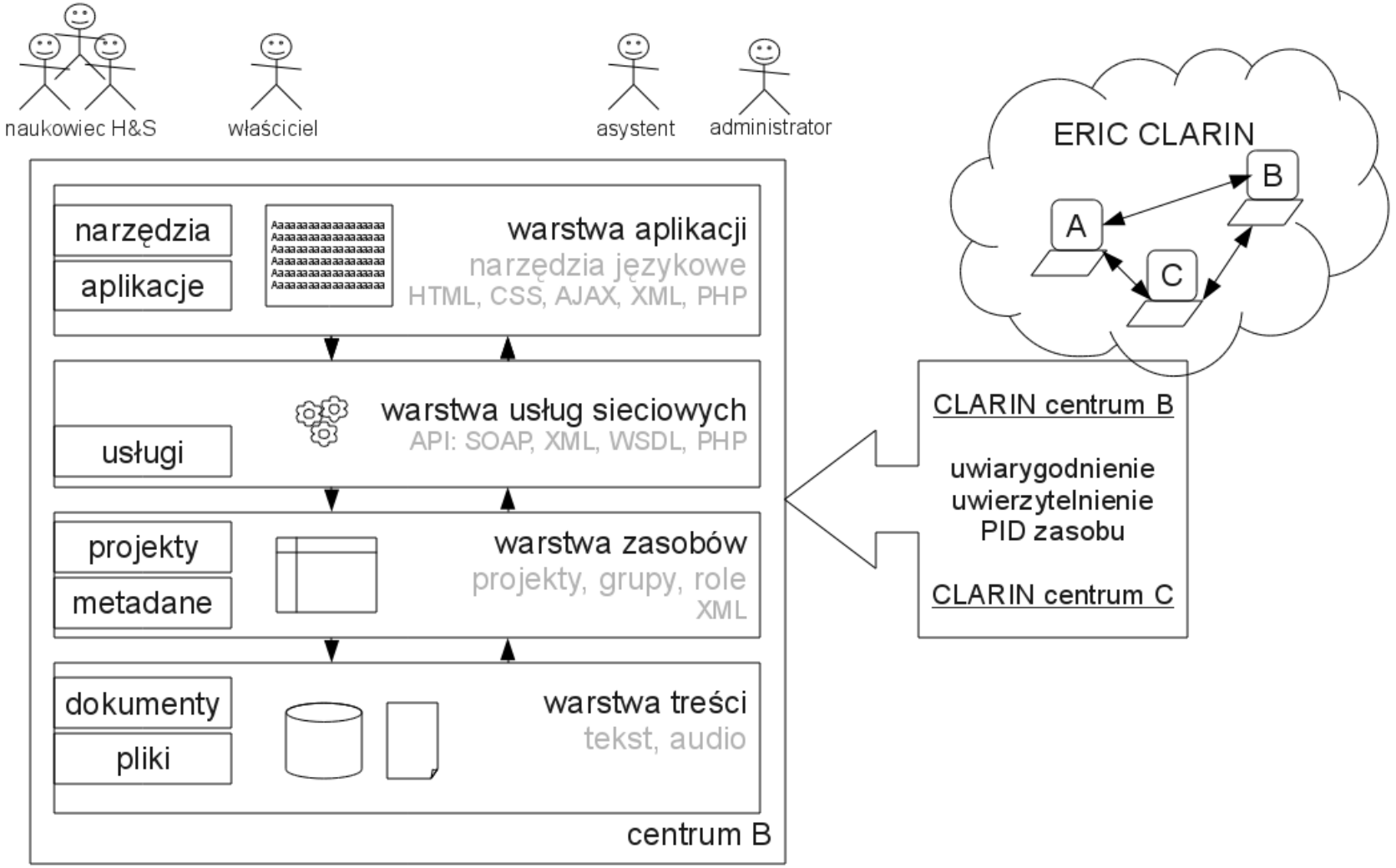
Autorzy filmu stawiają tezę, że pierwszym człowiekiem, który postawił nogę na wierzchołku Ziemi[Astro_Object] *był George Mallory*[Person].

Autorzy filmu[NP] stawiają tezę, że [*pierwszym człowiekiem, który postawił nogę na wierzchołku Ziemi*] *był George Mallory.*

Bariery w dostępie

- Fizyczna
  - narzędzia i zasoby nie są dostępne w sieci
- Informacyjna
  - brak opisu narzędzi i zasobów
  - brak katalogów i możliwości łatwego odnalezienia
- Technologiczna
  - brak standardów, możliwości łączenia elementów technologii
  - brak wspólnej platformy – różnorodność rozwiązań technologicznych
  - brak sprzętu o określonych parametrach
- Wiedzy
  - wymagane umiejętności programistyczne
  - wymagana wiedza z zakresu inżynierii języka naturalnego
- Prawna
  - licencje ograniczające dostęp i wykorzystanie
  - szczególnie w odniesieniu do korpusów

Infrastruktura językowa



Funkcje infrastruktury

- Odpowiedni system składowania (repozytoryjny)
  - trwałość danych (system archiwizacji)
  - jednoznaczny opis danych za pomocą *trwałych identyfikatorów* (Persistent Identifiers)
  - metadane o złożonej strukturze (CMDI)
  - zarządzanie metadanymi zgodnie z przyjętymi standardami (np. ISOcat, RELcat)
  - wirtualne kolekcje oparte na metadanych
- Rozproszona identyfikacja i autoryzacja użytkowników
  - oparta na federacjach narodowych
  - zasada jednego konta i jednego logowania
- Integracja zasobów i usług
  - w oparciu o usługi sieciowe (Web Services)
  - dostęp poprzez aplikacje sieciowe
  - brak konieczności ściągania i instalowania

Centra CLARIN

- Typ A – *centrum infrastrukturalne*
  - dostawca podstaw technologicznych i usług potrzebnych do podstawowego funkcjonowania sieci CLARIN
  - np. gromadzenia i agregacji metadanych, dostarczanie unikalnych identyfikatorów zasobów i narzędzi, itp.,
- Typ B – *centrum technologii językowych*
  - podstawowy składnik sieci
  - usługi, narzędzia, zasoby i aplikacje związane z przetwarzaniem języka naturalnego
- Typ C – *centrum metadanych*
  - umożliwiają automatczny dostęp do opisów zasobów (ale nie same zasoby)
- Typ K – centrum wiedzy
  - dostęp do wiedzy i ekspertów
  - wsparcie użytkowników CLARIN

Aplikacje – przykłady

- Ułatwienie dostępu
  - połączony katalog metadanych
  - federacyjne wyszukiwanie w korpusach tekstu i mowy
- Gromadzenie i zarządzanie danymi
  - tworzenie własnych kolekcji
  - rozszerzanie istniejących
  - wykorzystanie istniejących archiwów
- Rozszerzenie wyszukiwania w zasobach
  - automatyczna generacja metadanych w oparciu o narzędzia językowe
- Wydobywanie informacji i wiedzy
  - automatyczna generacja zestawień
  - analiza statystyczna oparta na faktach wydobytych z korpusu

CLARIN-PL

- Konsorcjum CLARIN-PL: polska część infrastruktury CLARIN
- Centrum Technologii Językowych CLARIN-PL
  - zlokalizowane na Politechnice Wrocławskiej
  - budowane w ramach Grupy Naukowej G4.19
  - zapewniające funkcje sieciowe infrastruktury CLARIN
  - udostępniające
    - repozytorium
    - zestaw wybranych aplikacji zbudowanych we współpracy z użytkownikami
    - wsparcie dla użytkowników – naukowców
- Korpusy
- Uzupełnienie brakujących elementów podstawowej technologii językowej dla języka polskiego
- Wybrane zasoby dwujęzyczne

CLARIN-PL

- Typowy schemat przetwarzania wypowiedzi
  1. Rozpoznanie struktury dokumentu i wydobycie tekstu
  2. Segmentacja: na tokeny, zdania oraz jednostki bardziej złożone
  3. Analiza morfologiczna
  4. Ujednoznacznianie morfo-syntaktyczne (tagowanie)
  5. Ujednoznacznianie sensu słów (znaczeń leksykalnych)
  6. Płytka analiza składniowa (płytki parsing) (opcjonalnie)
  7. Rozpoznawanie wyrażeń wielowyrazowych, w tym jednostek identyfikujących, np. nazw własnych.
  8. Rozpoznawanie związków w tekście, np. anafory, koreferencji, relacji semantycznych, sytuacji.
  9. Głęboka analiza składniowa (głęboki parsing)
  10. Głęboka analiza semantyczna *(częściowo)*
  11. Analiza pragmatyczna (w tym struktury dyskursu)

CLARIN-PL: wybrane zadania

- System długoterminowego przechowywania danych cyfrowych
- Korpusy: mowy, transkrypcji mowy, historyczny i dwujęzyczne
- Narzędzia do zaawansowanego przeszukiwania korpusów mowy i tekstu oraz wydobywania wiedzy lingwistycznej z korpusów
- Leksykalne zasoby semantyczne: bardzo duży dwujęzyczny wordnet, wyrażenia wielowyrazowe, nazwy własne oraz ramy walencyjne (struktury argumentowe)
- Płytkie i głębokie parsery semantyczne dla języka polskiego
- Wydobywanie informacji: rozpoznawanie nazw własnych, odniesień do czasu, anafory, relacji oraz sytuacji
- Narzędzia do automatycznego streszczania
- Narzędzia do wydobywania wiedzy z tekstu (Text Mining) ukierunkowane na zastosowania w naukach humanistycznych i społecznych – współpraca z użytkownikami

CLARIN-PL: projekt

- Okres: 2013-2015
- Partnerzy:
  - Politechnika Wrocławska, Instytut Informatyki (lider)
  - Instytut Podstaw Informatyki Polskiej Akademii Nauk
  - Instytut Slawistyki Polskiej Akademii Nauk
  - Polsko-Japońska Wyższa Szkoła Technik Komputerowych
  - Uniwersytet Łódzki
  - Uniwersytet Wrocławski

Dziękuję bardzo za uwagę