

Narzędzia i zasoby do przetwarzania mowy (korpusy, segmentacja, rozpoznawanie i analiza mowy)

Danijel Koržinek

Polsko-Japońska Akademia Technik Komputerowych

CLARIN-PL
Common Language Resources and Technology Infrastructure



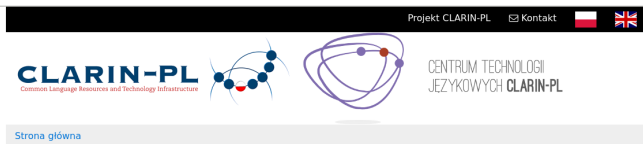
CENTRUM TECHNOLOGII
JEZYKOWYCH **CLARIN-PL**

12. kwietnia 2018 r., Poznań

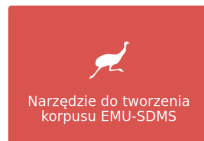
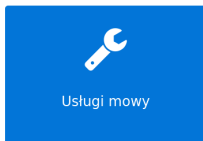
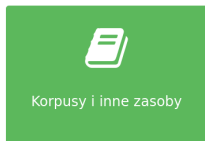
Strona mowa.clarin-pl.eu

Strona usług i narzędzi mowy

<http://mowa.clarin-pl.eu/>



Narzędzia i usługi do przetwarzania nagrań mowy



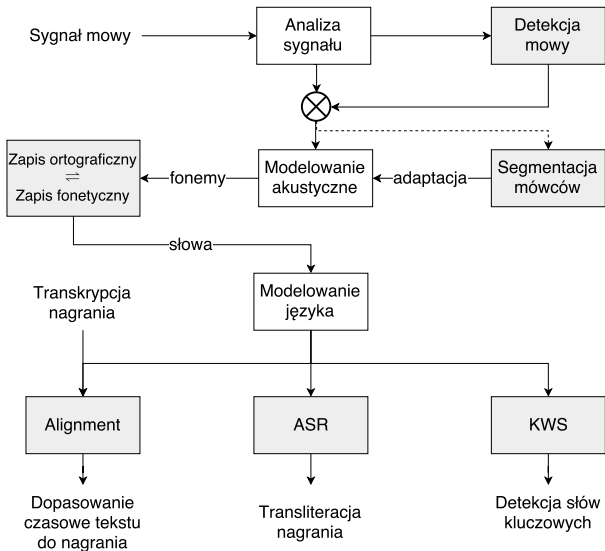
Korpus

- ▶ Korpus studyjny mowy czytanej
- ▶ Zapis: 16 kHz, 16-bit PCM
- ▶ Przeznaczony do podstawowego modelowania akustycznego mowy
- ▶ Rozmiary:
 - ▶ Długość: 56 godzin
 - ▶ Ilość mówców: 317
 - ▶ Ilość sesji: 554
 - ▶ Rozmiar sesji: 20 zdań + 10 słów bogatych fonetycznie
- ▶ Mniejszy korpus nagrany przez telefon komórkowy

Dystrybucja

- ▶ Wszystkie dane są dostępne na stronie:
<http://mowa.clarin-pl.eu/korpusy>
- ▶ System baseline dostępny na:
<https://github.com/danijel3/ClarinStudioKaldi>
- ▶ Wybrano licencję: CLARIN PUB+BY+INF+NORED
- ▶ Szczegółowe informacje na temat licencji:
<http://mowa.clarin-pl.eu/korpusy/LICENSE>

Technologie mowy



Usługi

<http://mowa.clarin-pl.eu/>



Usługi

- ▶ Konwersja tekstu na zapis fonetyczny (G2P)
- ▶ Normalizacja tekstu i audio
- ▶ Segmentacja (dopasowanie czasowe tekstu do nagrania)
- ▶ Transliteracja (rozpoznawanie mowy)
- ▶ Detekcja słów kluczowych
- ▶ Rozpoznawanie (diaryzacja) mówców
- ▶ Detekcja mowy (VAD)

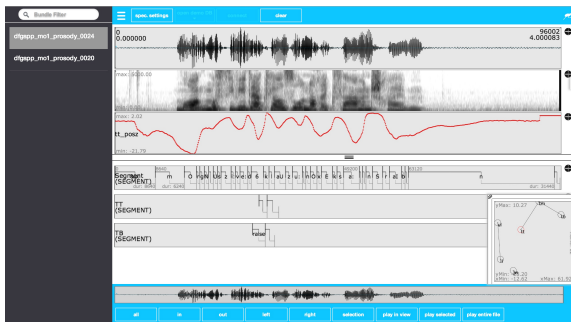
Korpusowa analiza mowy

Korpusowa analiza mowy

- ▶ Korpus mowy
 - ▶ nagranie + metadane
 - ▶ np. transkrypcja, mówcy, inne zjawiska, ...
 - ▶ wiele warstw opisu
 - ▶ opis numeryczny (na poziomie ramek/sygnалу)
 - ▶ np. formanty, energia, ...
- ▶ Badania wykorzystujące korpusy mowy:
 - ▶ fonetyka, lingwistyka, socjologia, psychologia, medycyna, ...

EMU-SDMS

- ▶ EMU Speech Database Management System
 - ▶ baza danych do korpusów mowy
- ▶ <http://ips-lmu.github.io/EMU.html>



Opis korpusu w formacie EMU

- ▶ “bundle” (paczka) - jedno nagranie i jego metadane
- ▶ sesja - grupa paczek (dowolnie dobrana)
- ▶ warstwy anotacji - opisujące zjawiska występujące w nagraniu
- ▶ hierarchie anotacji - pokazujące połączenia między warstwami
 - ▶ może być kilka
- ▶ “track data” - zawierające opisy na poziomie sygnału/ramek sygnału
- ▶ perspektywy - umożliwiające wybór wyświetlanych informacji
- ▶ wizualizacje - wyświetlane w osobnej ramce

Cechy bazy EMU

- ▶ Korpus jako katalog na dysku
- ▶ Annotacja w formacie JSON
- ▶ Biblioteka do ekstrakcji podstawowych cech (wrassp i format SSFF)
- ▶ Import/Export do TextGrid
- ▶ Aplikacja webowa wykorzystująca WebSockets
- ▶ Możliwość edycji korpusu przy użyciu aplikacji webowej
- ▶ Biblioteka do języka R do przeszukiwania i robienia zestawień statystycznych

Problemy korpusu EMU

- ▶ Zmusza do organizacji danych w standardowy sposób
- ▶ Daje wiele gotowych narzędzi
- ▶ Jak wgrać istniejące korpusy?
- ▶ Jak tworzyć nowe korpusy?
- ▶ Jak modyfikować strukturę korpusu?
- ▶ Jak dodawać kolejne rodzaje anotacji danych?
- ▶ Czy powinniśmy ufać narzędziom automatycznym?
 - ▶ <https://github.com/drammock/praat-semiauto>

Tworzenie korpusu EMU przy użyciu strony

1. Wgrywamy plik audio
 - ▶ Plik jest normalizowany do jednolitej postaci
2. Wgrywamy plik z transkrypcją, lub ...
 - ▶ Tekst jest normalizowany do postaci, którą można łatwo transkrybować
3. ... używamy automatycznego rozpoznawania mowy
 - ▶ Ręcznie poprawimy błędy procesu rozpoznawania
4. Uruchamiamy automatyczną segmentację
 - ▶ Poprawiamy błędy w segmentacji
5. Ściągamy korpus w postaci EMU
 - ▶ Wczytujemy do środowiska R i wykonujemy dalsze badania

Język R

- ▶ Co jest potrzebne:
 - ▶ Interpreter i IDE
 - ▶ np. RStudio
 - ▶ Biblioteki: emuR i wrassp
- ▶ Przykłady:
 - ▶ https://daniel3.github.io/emuR_notebooks
- ▶ Jak zacząć:
 - ▶ Kursy
 - ▶ <https://academy.vertabelo.com/course/introduction-to-r>
 - ▶ <https://labmasters.pl/kursy-otwarte/r/r-1/>
 - ▶ <http://goalkicker.com/RBook/>
 - ▶ Dla tych, którzy preferują Python: rpy2

Kontakt

Zapraszam do kontaktu:

- ▶ Danijel Koržinek - danijel@pja.edu.pl