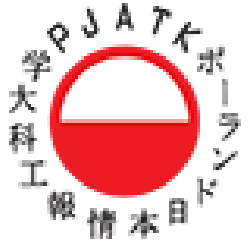
Korpusowa analiza mowy w języku R

Danijel Koržinek

Polsko-Japońska Akademia Technik Komputerowych



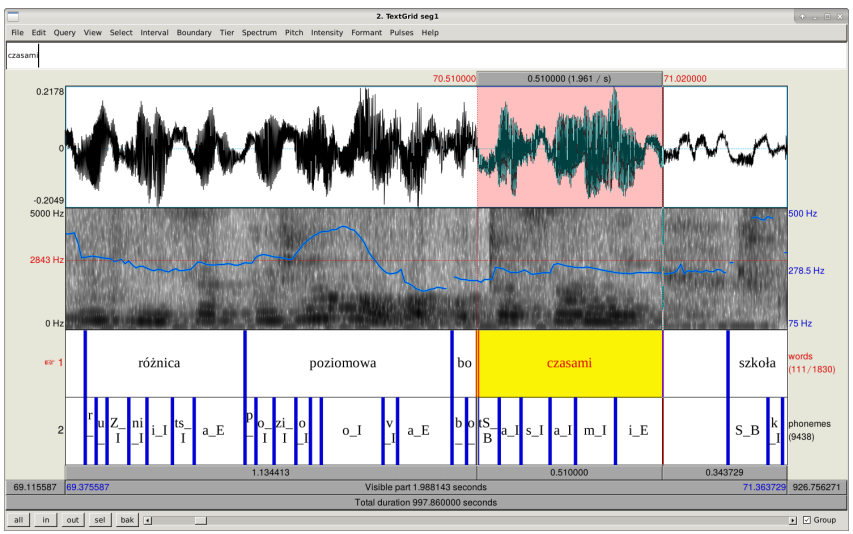
6. marca 2018 r., Warszawa

O mnie

- Typowy informatyk-programista
  - Basic, Pascal, C/C++, Java, Python
- Projekt Clarin-PL
  - http://www.clarin-pl.eu/
  - Ogólnoeuropejska infrastruktura naukowa
  - Umożliwia badaczom z dziedziny nauk humanistycznych i społecznych wygodną pracę z dużymi zbiorami danych
- Technologie i zasoby mowy
  - korpusy, rozpoznawanie mowy, segmentacja mowy, rozpoznawanie mówców, ...

Korpusowa analiza mowy

- Korpus mowy
  - nagranie + metadane
  - np. transkrypcja, mówcy, inne zjawiska, ...
  - wiele warstw opisu
  - opis numeryczny (na poziomie ramek/sygnału)
  - np. formanty, energia, ...
- Badania wykorzystujące korpusy mowy:
  - fonetyka, lingwistyka, socjologia, psychologia, medycyna, ...
- Jedno z bardziej popularnych narzędzi:
  - http://www.praat.org/

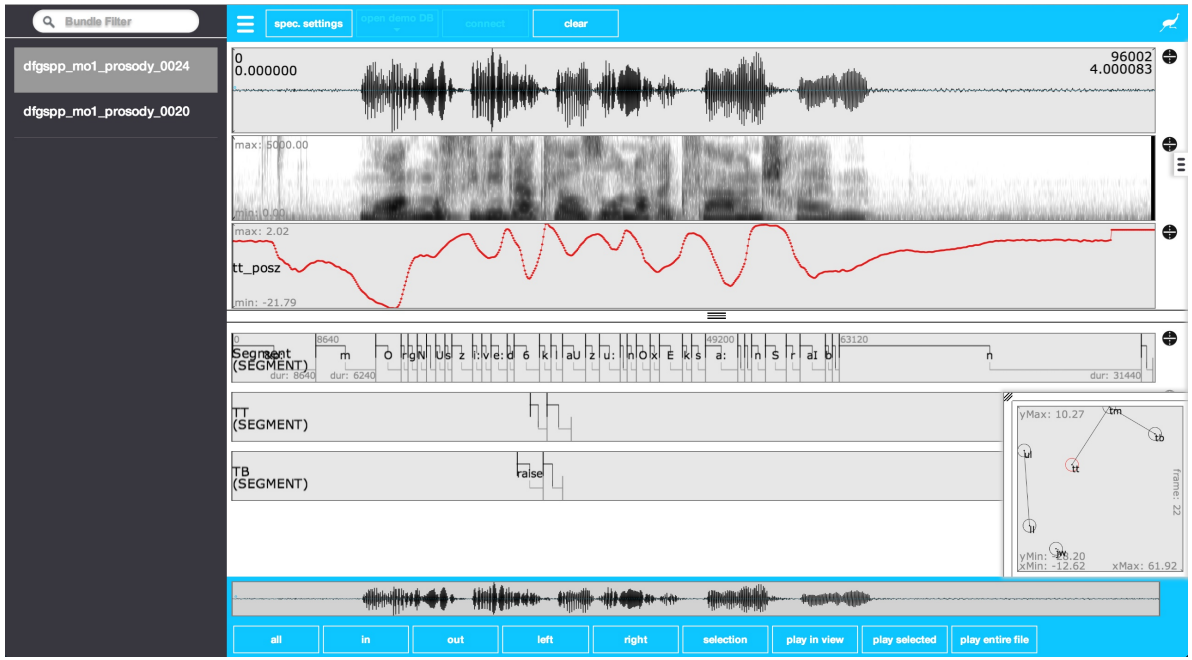


Narzędzia do automatycznej analizy

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

EMU-SDMS

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



Opis korpusu w formacie EMU

- Przykładowy korpus:
  - http://mowa.clarin-pl.eu/korpusy
- Podstawowe pojęcia:
  - “bundle” (paczka) - jedno nagranie i jego metadane
  - sesja - grupa paczek (dowolnie dobrana)
  - warstwy annotacji - opisujące zjawiska występujące w nagraniu
  - hierarchie annotacji - 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
  - https://danijel3.github.io/emuR_notebooks/

Dyskusja

- Czy emuR jest odpowiednim narzędziem dla polskich humanistów?
- Jak się polscy humaniści mogą nauczyć R do własnych zastosowań?
- Jak powinna wyglądać integracja zewnętrznych narzędzi z środowiskiem statystycznym?
- Jak powinny wyglądać narzędzia do tworzenia korpusów?
- Jakie inne technologie można zastosować w tym kontekście?
  - np. shiny?
- Zapraszam do kontaktu w razie pytań i sugestii:
  - Danijel Koržinek - danijel@pja.edu.pl