



18.05.2021

#PolskaHumanistykaCyfrowa || Automatyczna analiza fleksyjna tekstów polskich z lat 1830-1918 z uwzględnieniem zmian w odmianie i pisowni

Instytucje: Uniwersytet Warszawski (Instytucja kierująca)

Celem projektu jest stworzenie analizatora fleksyjnego dla polszczyzny drugiej połowy doby nowopolskiej, co jest wstępnym krokiem dla komputerowej analizy tekstów dawnych. Analizator ten będzie rozpoznawać formy z zachowaniem osobliwości odmiany i pisowni lat 1830-1918. Jego struktura zostanie tak zaprojektowana, by mógł on być podstawą dalszych rozszerzeń o stany wcześniejsze (przede wszystkim o pierwszą połowę doby nowopolskiej). Jednocześnie w ramach projektu, obok analizatora morfologicznego, powstanie nieduży (1 mln segmentów), oznakowany korpus tekstów z lat 1830-1918. Zasób ten zostanie udostępniony wraz z przeszukiwarką według kryteriów fleksyjnych i metatekstowych (periodyzacja próbek, ich przyporządkowanie do podzbiorów stylów funkcjonalnych, składniki opisu bibliograficznego).

Zasadnicze cele projektu to:

1. opis systemowych zmian w zakresie odmiany polszczyzny pisanej w latach 1830-1918;
2. stworzenie słownika fleksyjnego ukazującego ewolucję odmiany;
3. stworzenie małego korpusu gronowego (~1M segmentów, 1000 próbek) o równym pokryciu czasowym, z podziałem na podkorpusy stylów funkcjonalnych.

Drugi cel wiąże się z opracowaniem koncepcji leksykograficznej opisu zmian fleksyjnych i pisownianych w (elektronicznym) słowniku gramatycznym. Słownik ten będzie podstawą analizatora morfologicznego, zawierającego oprócz wzorców funkcjonujących do dziś paradygmaty funkcjonujące w badanym okresie, a współcześnie wygasłe.

Ponadto za cel postawiono stworzenie niewielkiego (1 mln segmentów) korpusu języka polskiego lat 1830-1918. Korpus ten będzie wykorzystany do identyfikacji ciągów nieznanymi analizatorowi oraz do testowania wzorców dodawanych. Korpus zostanie oznakowany i udostępniony (z możliwością przeszukiwania według zróżnicowanych kryteriów fleksyjnych i metatekstowych).

Punktem wyjścia prac będzie zgromadzenie korpusu o długości 1 mln, na który będzie się składać 1000 próbek o długości ok. 1000 segmentów

(priorytetem jest umieszczenie w próbce pełnych wypowiedzeń, nie zaś dokładna liczba słów). Próbki te będą zachowywać zróżnicowanie stylistyczne stosowane w *Słowniku frekwencyjnym polszczyzny współczesnej*. *

Institucja realizująca projekt:

- Instytut Języka Polskiego Uniwersytetu Warszawskiego

AUTOMATYCZNA ANALIZA FLEKSYJNA POLSZCZYZNY X

KLASY O PROJEKCIE PLIKI DO POBRANIA ZESPÓŁ PUBLIKACJE I WY

Analizator morfologiczny Morfeusz XIX-wieczny

Na tej stronie prezentujemy demonstracyjną wersję analizatora Morfeusz 2, którego słownik dostosowano do potrzeb analizy tekstów z lat 1830-1918. Słownik fleksyjny analizatora opracowano w latach 2013-2017 w ramach grantu Automatyczna analiza fleksyjna tekstów polskich z lat 1830-1918 z uwzględnieniem zmian w odmianie i pisowni (DEC-2012/07/B/H52/00570). Podstawą jest Słownik gramatyczny języka polskiego.

Podaj tekst:
[tłustej kompleksy, krótkiego oddechu]

Analizuj

II maja 2017

analizator Morfeusz2 ze słownikiem 1830-1918 on-line

Na stronie <http://sgjp.pl/morfeusz/demo-f19> można korzystać z analizatora Morfeusz wyposażonego w słownik polszczyzny 1830-1918. [...]

Projekt realizowany w ramach konkursu UMO 2013-2017 ze środków Ministerstwa Edukacji i Nauki (UMO DEC-2012/07/B/H52/00570)



II maja 2017

korpus f19 dostępny przez wyszukiwarkę Poliqarp/marasca

/f19/). [...]

Korpus polszczyzny 1830-1918 na stronie Szukaj w słownikach (<https://szukajwsownikach.uw.edu.pl>)



```
morfeusz_analyzer_version 1.9.2
setting dictionary search path to
loading dictionary morfeusz-f19
tłustej kompleksy krótkiego oddechu
1. tłustej, tłusty, adj sg dat f pos, ...
1. tłustej, tłusty, adj sg gen f pos, ...
1. tłustej, tłusty, adj sg loc f pos, ...
2. kompleksy, kompleksa, subst pl gen f, pospolita, 1835
2. kompleksy, kompleksa, subst sg dat f, pospolita, 1835
2. kompleksy, kompleksa, subst sg gen f, pospolita, 1835
2. kompleksy, kompleksa, subst sg loc f, pospolita, 1835]
3. krótkiego, krótki, adj sg acc m1 m2 pos, ..., 1834
3. krótkiego, krótki, adj sg gen m1 m2 m3 n1 n2 pos, ..., 1834
4. oddechu, oddech, subst sg gen m3, pospolita, ...
4. oddechu, oddech, subst sg loc m3, pospolita, ...
4. oddechu, oddech, subst sg voc m3, pospolita, ...
```

30 czerwca 2017

analizator Morfeusz z leksykonem f19

Statystyki nierozpoznanych segmentów przed modyfikacją analizatora i po niej. [...]

AKTUALNOŚCI

analizator Morfeusz z leksykonem f19
analizator Morfeusz z leksykonem f19
1830-1918 on-line
korpus f19 dostępny przez wyszukiwarkę Poliqarp/marasca
korpus, wersja zakodowana
lista haseł Słownika

korpus f19.0 beta
MultiDimensional Sorting

16 stycznia 2017

#PolskaHumanistykaCyfrowa - katalog humanistycznych projektów cyfrowych w Polsce. Celem powstania katalogu jest prezentacja różnorodnych projektów z zakresu szeroko pojętej humanistyki cyfrowej w Polsce. Katalog powstał w oparciu o ankietę przeprowadzoną wśród instytucji oraz osób prowadzących projekty, w których komponent cyfrowy stanowi kluczowy element pozyskania, analizy bądź prezentacji danych.

**informacje zaczerpnięte ze strony internetowej projektu*

Linki:

- <http://www.f19.uw.edu.pl/>
- <http://bit.ly/polskaHC>

Słowa kluczowe: fleksja, diachronia, polska humanistyka cyfrowa, korpus, ortografia, humanistyka cyfrowa, XIX wiek