

Informacje o korpusach

Czyli czym będziemy się zajmować na zajęciach

Bartosz Maćkiewicz

Instytut Filozofii, Uniwersytet Warszawski

18 lutego 2019

Jakie korpusy będziemy omawiać?

- ▶ Narodowy Korpus Języka Polskiego
- ▶ British National Corpus
- ▶ Corpus of Contemporary American English

Z jakich narzędzi będziemy korzystać?

- ▶ wyszukiwarki korpusowej **Pelcra** oraz kolokatora, która umożliwia pracę z NKJP;
- ▶ wyszukiwarki korpusowej **Poliqarp**, która umożliwia pracę z NKJP;
- ▶ multiwyszukiwarki i kolokatora **corpus.byu.edu** umożliwiającej pracę z COCA, BNC i wieloma innymi dostępnymi korpusami
- ▶ programu do zarządzania i przetwarzania korpusów **SketchEngine**

Jak tworzone są korpusy?

- ▶ próba vs całe teksty
- ▶ zróżnicowanie vs różnorodność
- ▶ anotacja
- ▶ oportunizm

Struktura tekstowa korpusów

reprezentatywność vs zrównoważenie

- ▶ **reprezentatywność** - odnoszenie się do jakiejś rzeczywistości istniejącej poza korpusem
- ▶ **zrównoważenie** - dbałość o taką budowę korpusu, żeby żaden składnik na żadnym z poziomów nie dominował nad innym.

Struktura tekstowa w NKJP

Górski i Łaziński (2012) podają następujące proporcje tekstów:

- ▶ Publicystyka i krótkie wiadomości prasowe: 50
- ▶ Literatura piękna: 16
- ▶ Literatura faktu: 5,5
- ▶ Typ informacyjno-poradnikowy: 5,5
- ▶ Typ naukowo-dydaktyczny: 2,0
- ▶ Inne teksty pisane: 3,0
- ▶ Książka niebeletrystyczna nieklasyfikowana: 1,0
- ▶ Teksty konwersacyjne, mówione medialne i quasi-mówione razem: 10,0
- ▶ Teksty internetowe statyczne i dynamiczne razem: 7,0

Jakie informacje są dostępne w korpusach?

metadane

„Aby zdjąć ze mnie ten straszny obowiązek, ten rozkaz piekielny, ksiądz zabije innego człowieka”

- ▶ **autor:** Jarosław Iwaszkiewicz
- ▶ **tytuł:** Brzezina i inne opowiadania Kościół w Skaryszewie
- ▶ **źródło:** Brzezina i inne opowiadania Kościół w Skaryszewie
- ▶ **ISBN:** 9788307030838
- ▶ **rok publikacji:** 2006
- ▶ **wydawca:** Czytelnik
- ▶ **miejsce publikacji:** Warszawa
- ▶ **typ:** literatura piękna
- ▶ **kanal:** książka

Jakie informacje są dostępne w korpusach?

anotacje

„Kto miał wg Ciebie obowiązek informowania opinii publicznej o tej sprawie?”

[informować:ger:sg:gen:n:imperf:aff]

- ▶ forma bazowa tego słowa to "informować"(lemat)
- ▶ klasa gramatyczna tego słowa to rzeczownik odczasownikowy (ger)
- ▶ jest to rzeczownik w liczbie pojedynczej (sg)
- ▶ przypadek tego rzeczownika to dopełniacz (gen)
- ▶ rodzaj tego rzeczownika to rodzaj nijaki (n)
- ▶ czasownik, od którego derywowana jest ta forma jest czasownikiem niedokonanym (imperf)
- ▶ jest to forma niezanegowana (czyli nie jest to np. /niepoinformowanie/) (aff)