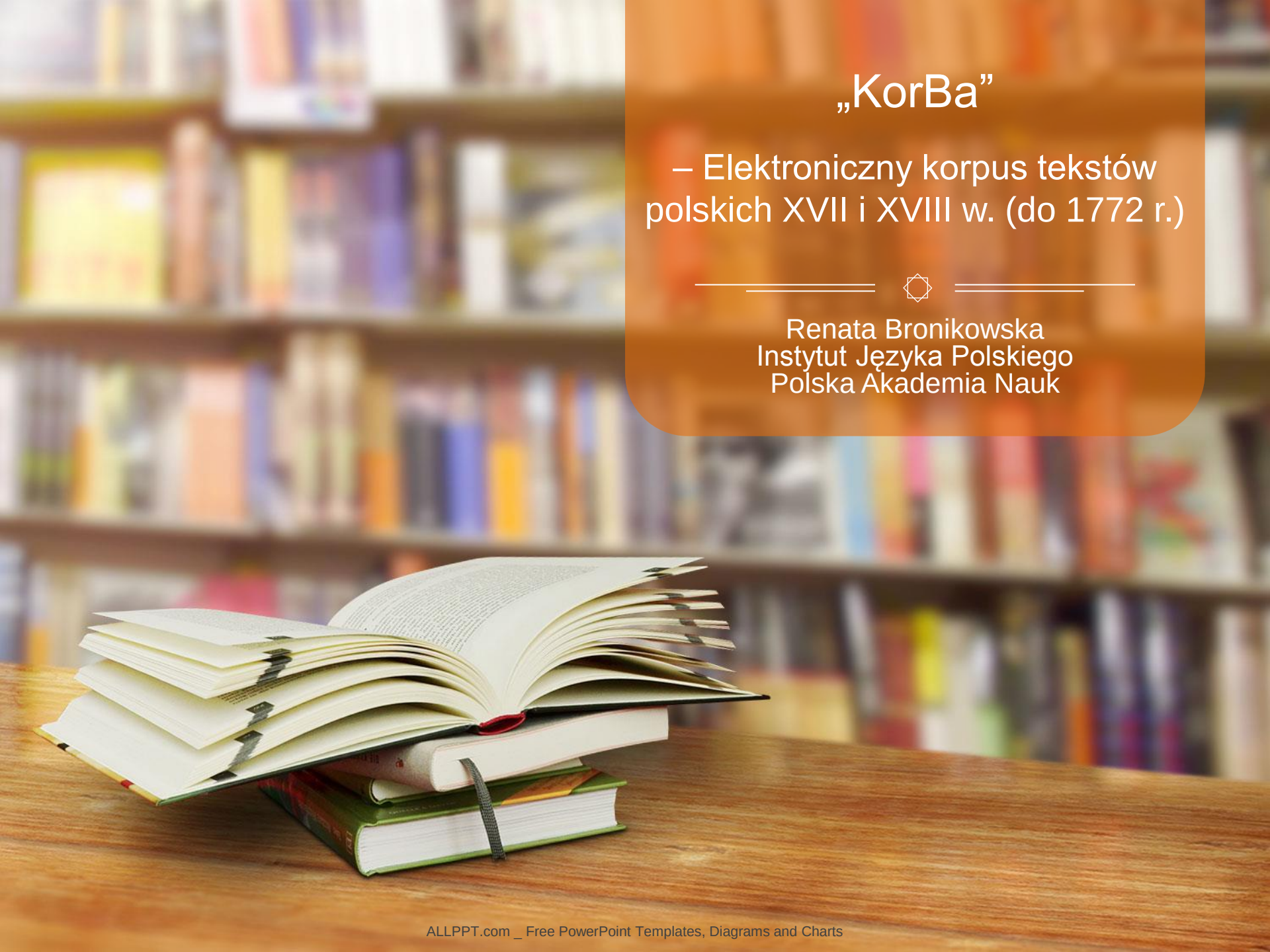
„KorBa”

– Elektroniczny korpus tekstów polskich XVII i XVIII w. (do 1772 r.)

Renata Bronikowska
Instytut Języka Polskiego
Polska Akademia Nauk



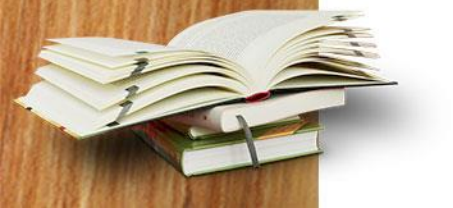
PODSTAWOWE INFORMACJE O PROJEKCIE

- Kryptonim: KORBA = KORpus BArokowy
- Cel: stworzenie obszernego (12 milionów segmentów) korpusu polskich tekstów XVII- i XVIII-wiecznych (do 1772 r.)
- Czas trwania: 2013-2018
- Jednostka koordynująca: Instytut Języka Polskiego Polskiej Akademii Nauk
- Współpraca: Instytut Podstaw Informatyki Polskiej Akademii Nauk
- Kierownik: Włodzimierz Gruszczyński
- Finansowanie: Narodowy Program Rozwoju Humanistyki (numer projektu 0036/NPRH2/H11/81/2012)

KorBa

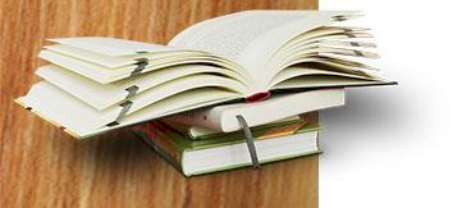
ZASTOSOWANIE

- Stworzenie historycznego podkorpusu Narodowego Korpusu Języka Polskiego (http://nkjp.pl).
- Przyspieszenie prac nad *Elektronicznym słownikiem języka polskiego XVII i XVIII w.* (http://sxvii.pl).
- Pozyskane dane posłużą do opracowania diachronicznego modelu fleksji polskiej.



OBECNY ETAP PRAC NAD KORPUSEM

- 718 tekstów o łącznej objętości 10 835 754 słów (ponad 12 mln segmentów w rozumieniu NKJP)
- Wszystkie teksty transliterowane oraz oznakowane strukturalnie i językowo, zapisane w formacie TEI XML
- Stworzony tagset barokowy
- Rozpoczęta ręczna anotacja morfosyntaktyczna 0,5-milionowego podkorpusu
- Nakładka na wyszukiwarkę Poliqarp 2, ułatwiająca przeszukiwanie korpusu



KorBa

METADANE

- Dane bibliograficzne
- Informacja o podstawie źródłowej:
  - Starodruki i rękopisy
  - Wydania XIX-, XX, XXI-wieczne
- Charakterystyka stylistyczno-genologiczna:
  - Rodzaj
  - Gatunek
  - Tematyka
  - Mowa wierszowana/ niewierszowana
  - Poetyka żartu
- Charakterystyka chronologiczna – podział na podokresy:
  - 1601-1650
  - 1651-1700
  - 1701-1750
  - 1751-1772
- Charakterystyka geograficzna – podział na regiony:
  - Małopolska
  - Wielkopolska
  - Mazowsze
  - Pomorze i Prusy
  - Ziemie Wlk. Ks. Lit.
  - Ziemie Ruskie
  - Śląsk

KORBA — wczytywacz tekstów

renata (Redaktor)

Strona główna
Nowy tekst

Wyszukiwarka
Raporty

Pracuj jako:
skryptor
korektor

Zgłoś błąd
Zmień hasło
Wyloguj

Wczytanych tekstów: , zawierających w sumie (wg TEI) słów.

Zakończonych tekstów: , zawierających w sumie (wg TEI) słów.

Powrót do strony głównej

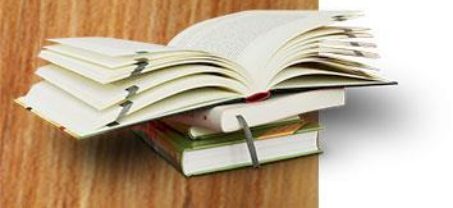
Edytuj metadane

Id: *	OvŻebrMet
Tytuł: *	Metamorphoseon
Autor: *	Publius Ovidius Naso ☐ Anonimowy
Tłumacz:	Jakub Żebrowski ☐ Anonimowy
Drukarnia:	Franciszek Cezary
Rok:	1636
Rok wydania niepewny (wydano nie wcześniej, niż podany rok):	☐
Typ mowy: *	wierszowana
Rodzaj: *	epika
Gatunek:	☐ pieśni ☐ fraszki i epigramaty ☐ epitafia ☐ satyry ☐ sielanki ☐ kazania ☐ pisma polityczne ☐ polemiki religijne ☐ mowy okolicznościowe ☐ traktaty ☐ dialogi ☐ pamiętniki ☐ kroniki ☐ relacje ☐ opisy podróży ☐ herbarze ☐ akta sejmikowe ☐ wilkierze ☐ księgi sądowe ☐ inwentarze ☐ rejestry ☐ diariusze sejmowe ☐ rozmówki do nauki języka ☐ podręcznik ☐ przysłowia ☐ kalendarze ☐ przewodniki ☐ żywoty świętych ☑ poematy epickie ☐ przypowieści, specula (zwierciadła) ☐ modlitwy ☐ panegiryk ☐ lamenty ☐ emblematy ☐ przywileje ☐ konstytucje sejmowe ☐ bajki ☐ księgi liturgiczne
Tematyka:	☐ alchemia ☐ anatomia ☐ architektura ☐ astrologia ☐ astronomia ☐ biologia ☐ botanika ☐ budownictwo ☐ chemia ☐ egzotyka ☐ ekonomia ☐ filozofia ☐ fizyka ☐ geografia ☐ gospodarstwo ☐ gramatyka ☐ górnictwo ☐ historia ☐ hutnictwo ☐ języki ☐ kulinaria ☐ matematyka ☐ medycyna ☑ mitologia ☐ miłość ☐ muzyka ☐ myślistwo ☐ obyczajowość ☐ poetyka ☐ polityka ☐ prawo ☐ religia ☐ retoryka ☐ wojskowość ☐ zielarstwo ☐ zoologia ☐ żeglarstwo
Poetyka żartu:	☐
Miejsce: *	Kraków
Region: *	Małopolska
Wydanie uwspółcześnione:	☐
Pochodzi z korpusu IMPACT:	☐
Tekst chroniony prawami autorskimi:	☐

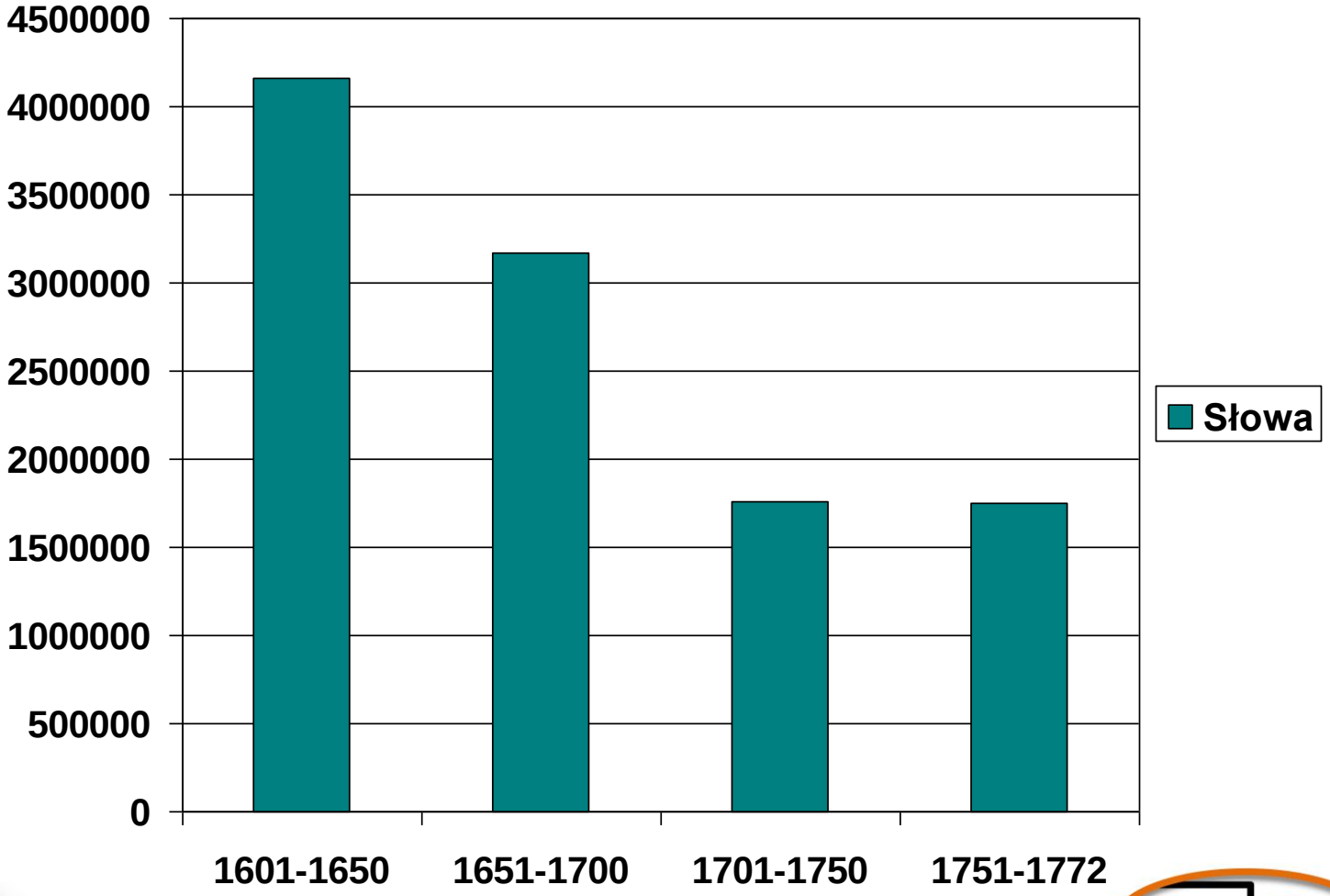
Aktualizuj

KORZYŚCI Z METADANYCH

- Możliwość ograniczenia przeszukiwania do pewnej grupy tekstów (np. teksty jednego autora, pochodzące z jednego regionu, powstałe w 1. poł. XVII w.)
- Dodatkowe informacje pozwalające na lepszą interpretację uzyskanych danych, np.
  - Postać wyrazu wymuszona przez wymagania rymu lub rytmu (← mowa wierszowana)
  - Ironiczne użycie wyrazu (← poetyka żartu)
  - Ostrożne podejście do danych pochodzących z wydań późniejszych, szczególnie XIX-wiecznych



CHRONOLOGICZNA REPREZENTACJA TEKSTÓW

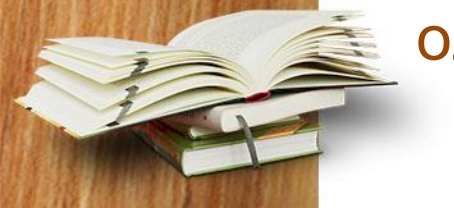


ZRÓŻNICOWANIE TEKSTÓW

- Podstawy źródłowe:
  - Starodruki i rękopisy: 64%
  - Wydania współczesne (XIX-, XX-, XXI-wieczne): 36%
- Typ literatury:
  - Literatura piękna (w tym Biblia): 26%
  - Pozostałe: 74%
- Obecność rymów:
  - Mowa niewierszowana: 76%
  - Mowa wierszowana: 21%
  - Mowa mieszana: 3%
- Poetyka żartu: 2%

TYPY ANOTACJI

- Anotacja strukturalna – oznaczanie wyodrębnionych fragmentów struktury tekstu (np. strona, rozdział, notka marginesowa), a także elementów pominiętych, np. ilustracja, dłuższy fragment w języku obcym;
- Anotacja językowa – oznaczanie obcojęzycznych fragmentów tekstu, np. łaciny;
- Anotacja morfosyntaktyczna – przypisanie informacji gramatycznej każdemu segmentowi; obecnie rozpoczęliśmy ręczne znakowanie 0,5-milionowej próbki, która będzie potem użyta do budowy tagera, służącego do automatycznego oznakowania pozostałej części korpusu.

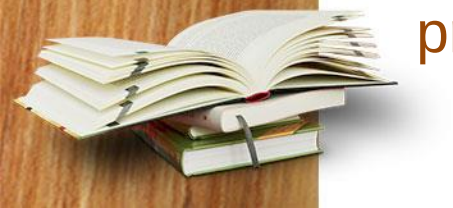


KORZYŚCI Z ANOTACJI

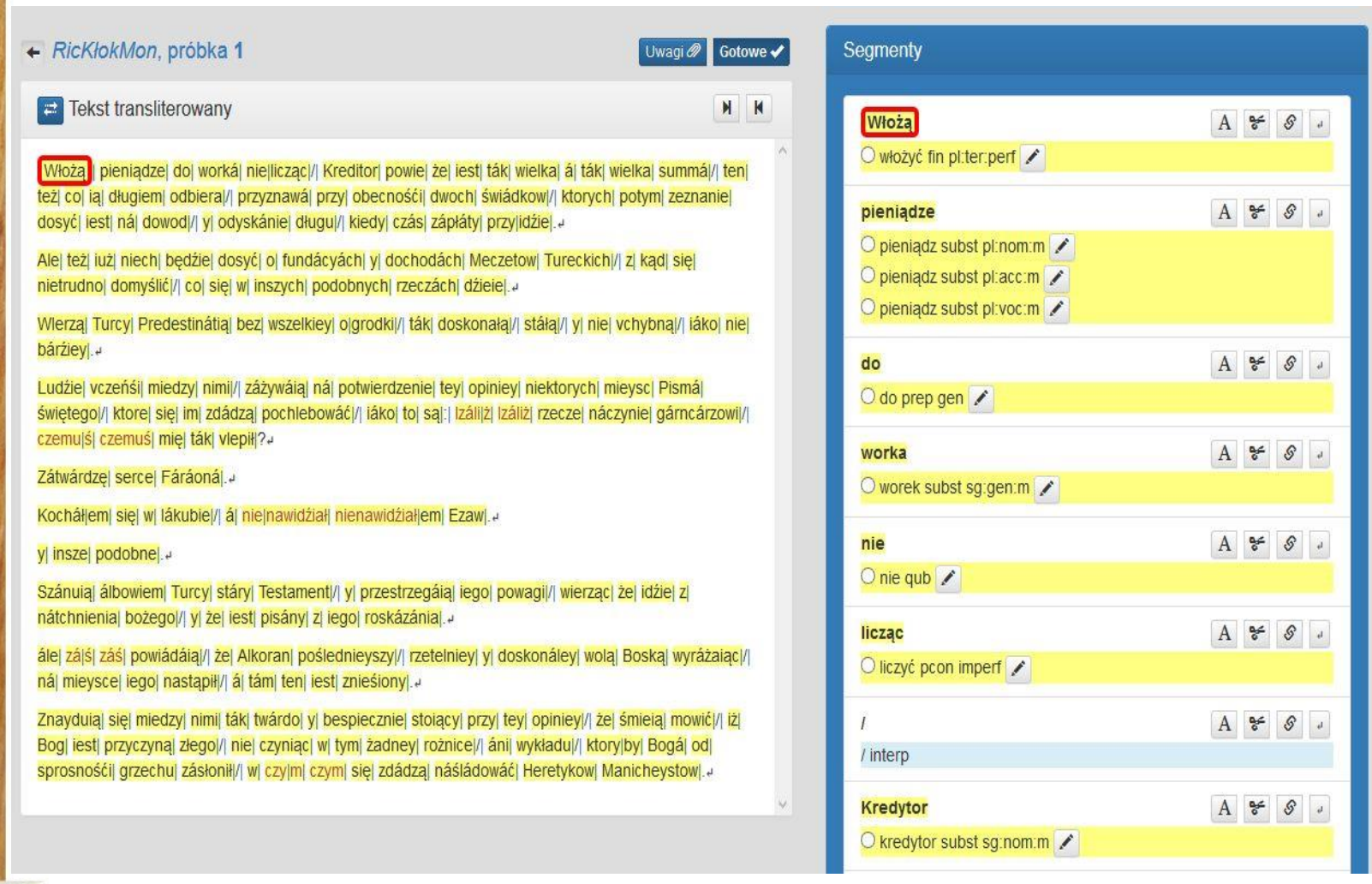
- Możliwość zastosowania zaawansowanej składni zapytań
- Teksty dostępne w wersji transliterowanej i transkrybowanej
- Za pomocą jednego zapytania można wyszukać formy zróżnicowane graficznie (np. *rada* – *rádá, panem* – *panẽ*)
- Dokładna lokalizacja wyszukanych wyrażeń w źródle – powiązanie każdego segmentu z numerem strony
- Fragmenty obcojęzyczne wyłączone z przeszukiwania, ale pojawiające się jako kontekst szukanego wyrażenia
- Pełniejsza informacja o kontekście, w jakim występuje szukane wyrażenie (np. strona tytułowa, żywa pagina)
- Dodatkowe informacje o zapisie tekstu: odczytanie niepewne, tekst nieczytelny, brak fragmentu tekstu

NARZĘDZIA

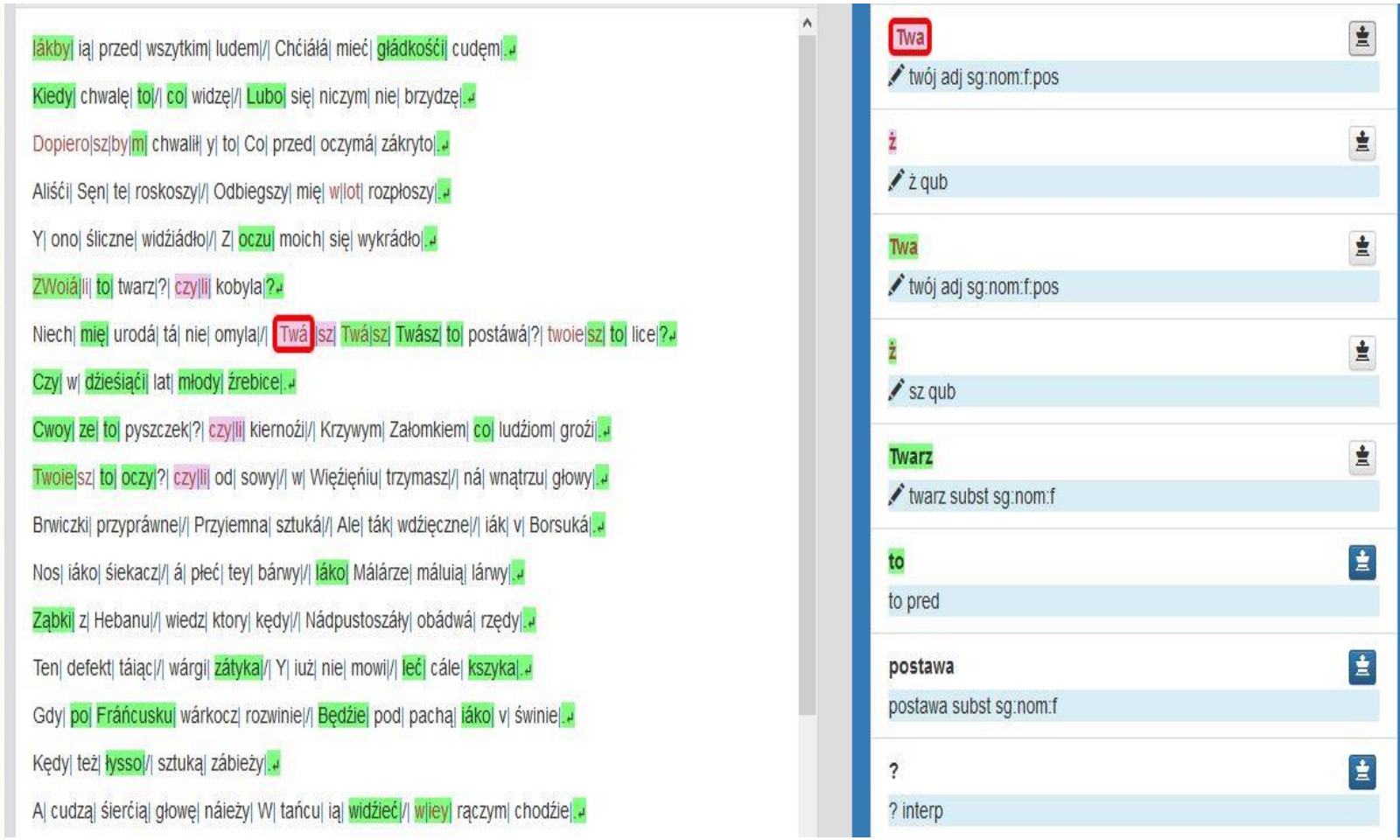
- Wczytywacz tekstów – konwertuje pliki wordowe na format xml zgodny z TEI, wykrywa błędy w anotacji strukturalnej i językowej, pokazuje statystyki korpusu;
- Konwerter – przekształca teksty transliterowane na transkrybowane (https://bitbucket.org/jsbien/pol);
- Anotatornia 2 – wspomaga ręczne znakowanie morfosyntaktyczne;
- Korbeusz – analizator morfologiczny, dostosowany do analizy tekstów XVII- i XVIII-wiecznych;
- Tager – służy do automatycznego znakowania morfosyntaktycznego tekstów;
- Poliqarp 2 – wyszukiwarka dostosowana do przeszukiwania korpusu barokowego.



PRACA ANOTATORA



PRACA SUPERANOTATORA



Dziękuję za uwagę