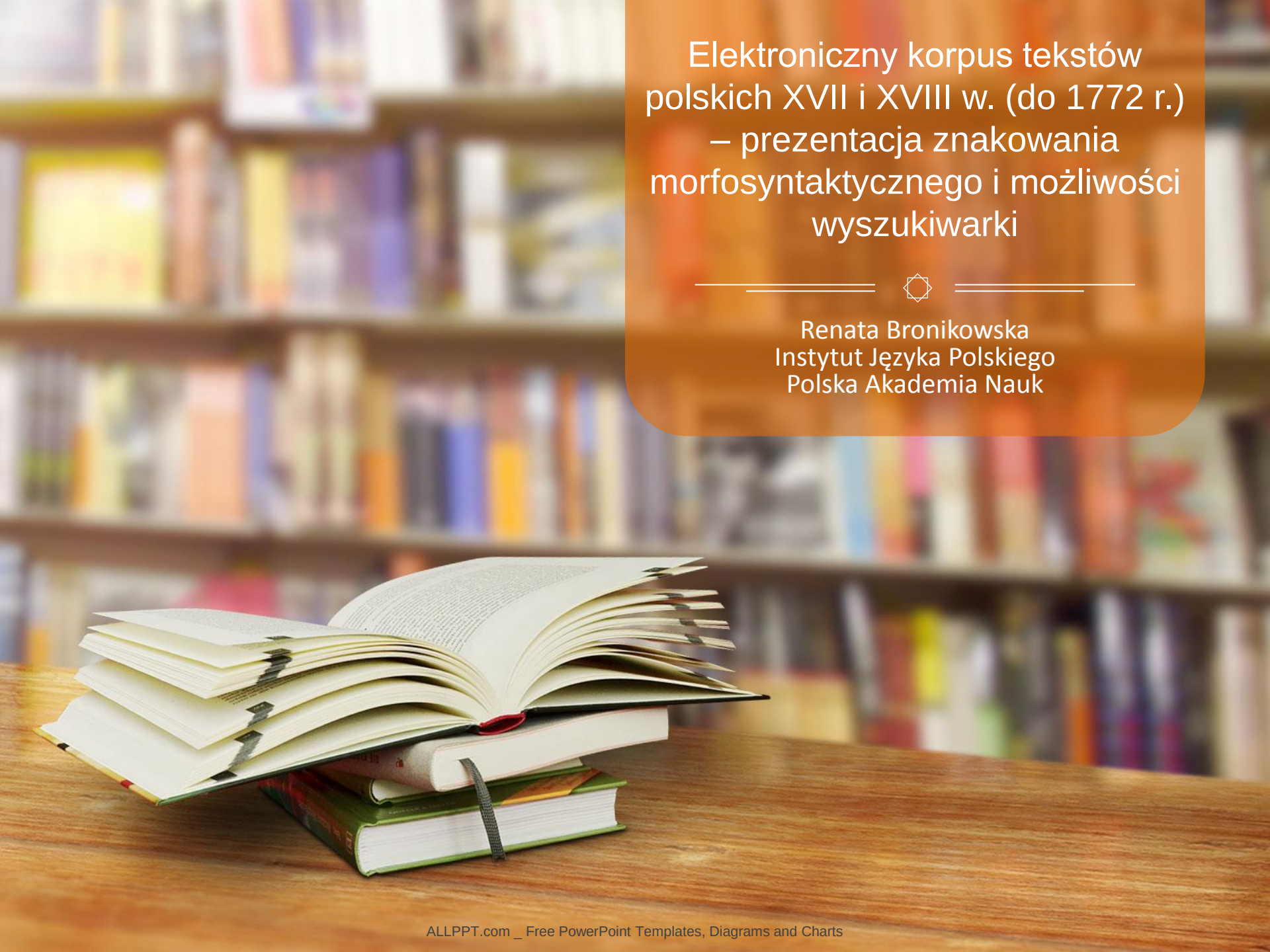
Elektroniczny korpus tekstów polskich XVII i XVIII w. (do 1772 r.) – prezentacja znakowania morfosyntaktycznego i możliwości wyszukiwarki

Renata Bronikowska
Instytut Języka Polskiego
Polska Akademia Nauk



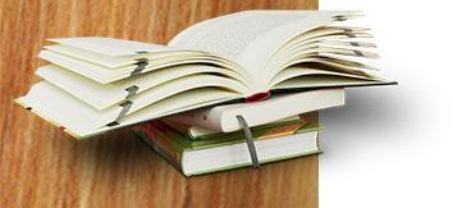
PODSTAWOWE INFORMACJE O PROJEKCIE

- 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

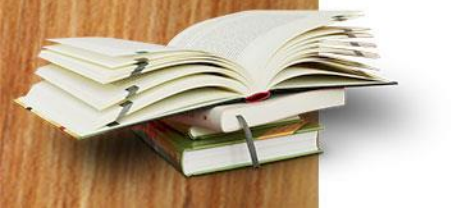
ZASTOSOWANIE

- 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

- 717 tekstów o łącznej objętości 10 808 728 słów
- Typy tekstów: starodruki, rękopisy, XIX-, XX- i XXI-wieczne wydania tekstów barokowych
- 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



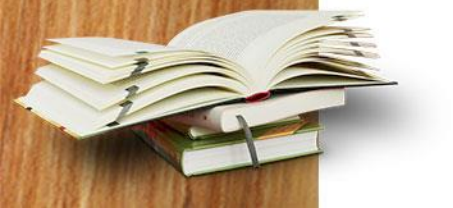
KorBa

PRZYGOTOWANIE PRÓBEK DO ANOTACJI

- Z każdego tekstu są automatycznie losowane próbki, z których każda zawiera ok. 200 segmentów.
- Liczba próbek pochodząca z każdego tekstu jest proporcjonalna do jego objętości.
- Próbki zostają przekonwertowane z postaci transliterowanej na transkrybowaną, tak aby ich zapis zbliżył się do zapisu współczesnego.

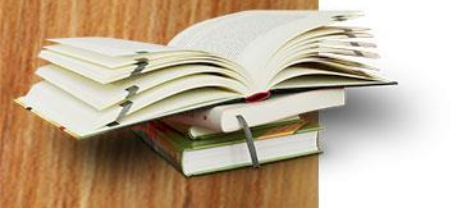
TRANSKRYBER

- Narzędzie stworzone przez Janusza Bienia i jego zespół na potrzeby projektu IMPACT (https://bitbucket.org/jsbien/pol).
- Składa się z zestawu reguł (obecnie ponad 3000), które określają sposób zamiany poszczególnych liter albo grupy liter w określonym kontekście.
- Np. rozpoczynająca słowo grupa *naiw* powinna zostać zamieniona na *najw* (np. *najwyższy*), z wyjątkiem sytuacji, kiedy następuje po niej litera *n* (np. *naiwny*).
- Oryginalny zestaw reguł został rozbudowany na potrzeby korpusu barokowego przez Emanuela Modrzejewskiego.



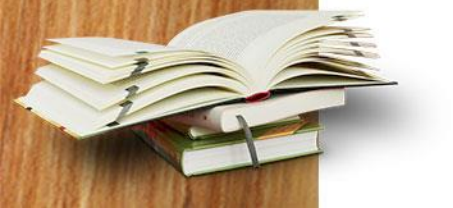
ANOTATORNIA 2

- Narzędzie wspomagające ręczną anotację morfosyntaktyczną autorstwa Doroty Komosińskiej.
- Umożliwia jednoczesną pracę nad próbką dwojga anotatorów oraz wprowadzanie poprawek przez superanotatora .
- Praca anotatorów polega na wyborze odpowiedniej interpretacji morfologicznej z kilku zaproponowanych przez analizator morfologiczny albo ręcznym wpisaniu własnej.



KORBEUSZ

- Analizator morfologiczny dostosowany do obsługi tekstów barokowych = zmodyfikowany Morfeusz 2 autorstwa Marcina Wolińskiego.
- Wykorzystuje do analizy różnego rodzaju dane:
  - pełne paradygmaty polskich leksemów zawarte w SGJP
  - niepełne paradygmaty leksemów XVII-XVIII-wiecznych pobrane ze słownika e-SXVII
  - automatycznie tworzone rekonstrukcje paradygmatów barokowych
- Zawiera także reguły segmentacyjne, które uwspółcześniają zapis barokowy, np. oddzielają przyimek od rzeczownika.



PRACA W ANOTATORNI

- http://test.anotatornia.nlp.ipipan.waw.pl

PRZESZUKIWANIE KORPUSU

- http://test.korba.edu.pl/