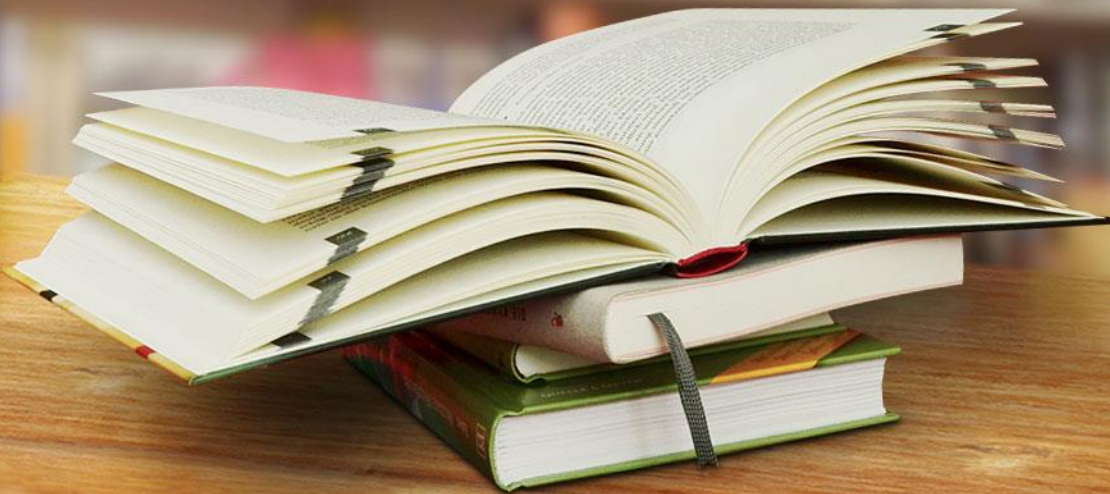


„KorBa”

– 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)



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



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



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

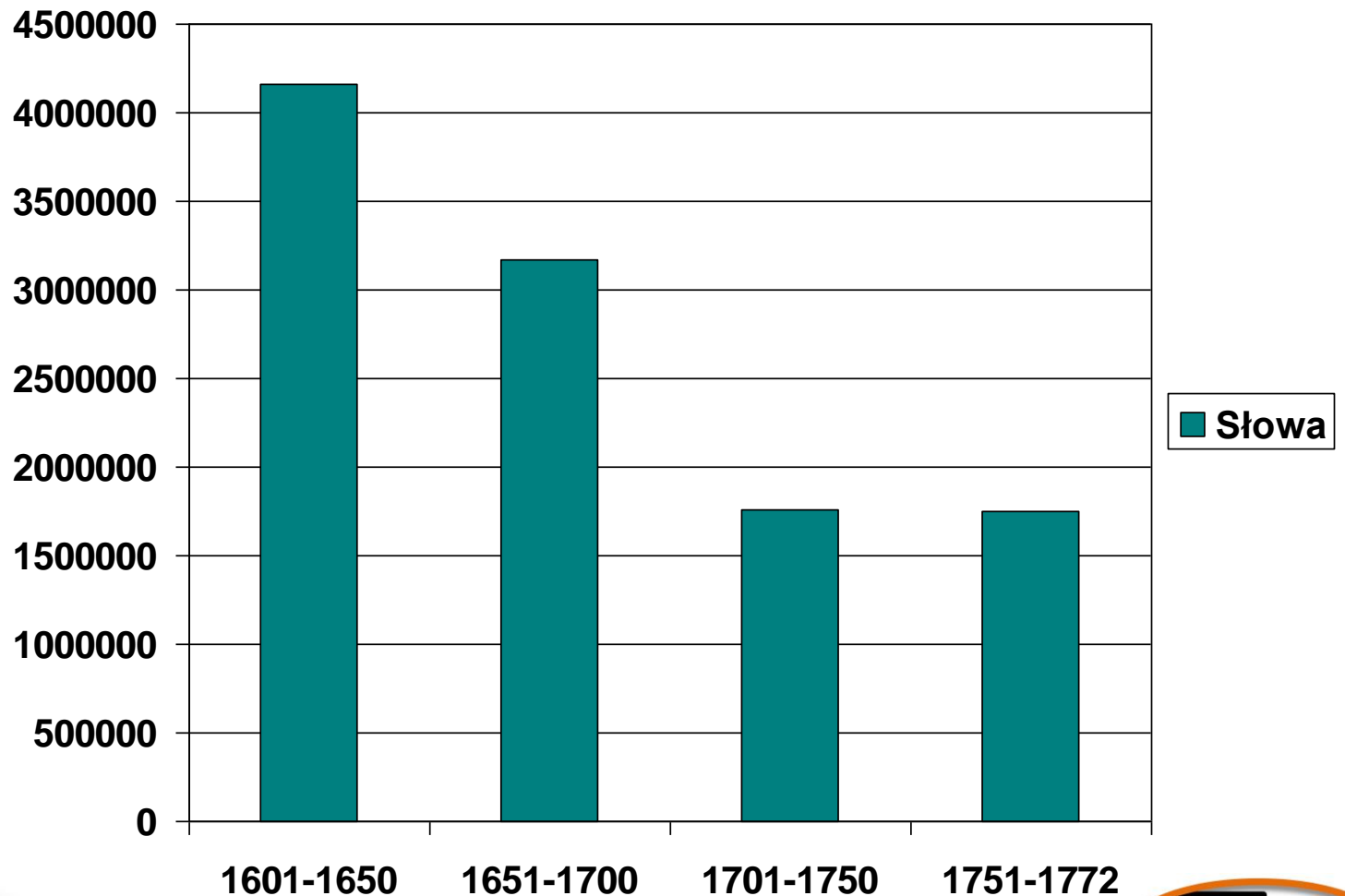
Id: *	<input type="text" value="OvŻebrMet"/>
Tytuł: *	<input type="text" value="Metamorphoseon"/>
Autor: *	<input type="text" value="Publius Ovidius Naso"/> <input type="checkbox"/> Anonimowy
Tłumacz:	<input type="text" value="Jakub Żebrowski"/> <input type="checkbox"/> Anonimowy
Drukarnia:	<input type="text"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

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* – *paně*)
- Dokładna lokalizacja wyszukanych wyrażen 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

← RicKłokMon, próbka 1 Uwagi Gotowe

Tekst transliterowany

Włoża pieniądze| do| worka| nie|licząc|/| Kreditor| powie| że| jest| tak| wielka| a| tak| wielka| summá|/| ten| też| co| ją| długiem| odbiera|/| przyznawá| przy| obecności| dwóch| świadków|/| których| potom| zeznanie| dosyć| jest| ná| dowod|/| y| odyskánie| długu|/| kiedy| czas| zápláty| przy|idzie|.↵

Ale| też| już| niech| będzie| dosyć| o| fundácii| y| dochódach| Meczetow| Tureckich|/| z| kąd| się| nietrudno| domyślić|/| co| się| w| inszych| podobnych| rzeczách| dzieie|.↵

Wierzą| Turcy| Predestináciá| bez| wszelkiej| o|grodkí|/| tak| doskonałą|/| stáją|/| y| nie| vchybną|/| iáko| nie| bárżey|.↵

Ludzie| vceńší| między| nimi|/| záżywáją| ná| potwierdzenie| tey| opiniey| niektórych| mieysc| Pismá| świętego|/| ktore| się| im| zdádzą| pochlebowác|/| iáko| to| są|/| Izálli|/| Izálli| rzecz| náczynie| gámcárowi|/| czemu|s| czemu|s| mié| tak| vlepi|?|.↵

Zátwárdę| serce| Fáráoná|.↵

Kochám|em| się| w| iákubie|/| á| nie|nawidziám| nienawidziám|em| Ezaw|.↵

y| insze| podobne|.↵

Szánuia| álbowiem| Turcy| stáry| Testament|/| y| przestrzegáją| iego| powagi|/| wierząc| że| idzie| z| nátnhienia| bożego|/| y| że| jest| pisány| z| iego| roskázániá|.↵

ále| zá|s| zá|s| powiádáją|/| że| Alkoran| poslednieysz|/| rzetelniey| y| doskonáley| wołá| Boską| wyráziác|/| ná| mieysce| iego| nastápi|/| á| tám| ten| jest| zniešiony|.↵

Znayduia| się| między| nimi| tak| twardo| y| bezpiecznie| stoiący| przy| tey| opiniey|/| że| śmieia| mowić|/| i| Bog| jest| przyczyná| złego|/| nie| czyniac| w| tym| żadney| różnice|/| áni| wykładu|/| który|by| Bogá| od| sprosności| grzechu| zášloni|/| w| czy|/m| czym| się| zdádzą| nášládowác| Heretykow| Manicheystow|.↵

Segmenty

Włoża A ✂ 🔗 ⌵

włożyć fin pl:ter:perf ✎

pieniądze A ✂ 🔗 ⌵

pieniądz subst pl:nom:m ✎

pieniądz subst pl:acc:m ✎

pieniądz subst pl:voc:m ✎

do A ✂ 🔗 ⌵

do prep gen ✎

worka A ✂ 🔗 ⌵

worek subst sg:gen:m ✎

nie A ✂ 🔗 ⌵

nie qub ✎

licząc A ✂ 🔗 ⌵

liczyć pcon imperf ✎

/ A ✂ 🔗 ⌵

/ interp

Kredytor A ✂ 🔗 ⌵

kredytor subst sg:nom:m ✎



PRACA SUPERANOTATORA

lákby| ią| przed| wszystkim| ludem|/| Chciał| mieć| gładkość| cudem|.↵
Kiedy| chwale| to|/| co| widzę|/| Lubo| się| niczym| nie| brzydę|.↵
Dopiero|sz|by|**m**| chwali| y| to| Co| przed| oczym| zakryto|.↵
Aliści| Sę| te| roskosz|/| Odbieg|szy| mi| w|lot| rozpłosz|.↵
Y| ono| ślicz|nej| widzi|adło|/| Z| **oczu**| moich| się| wykr|adło|.↵
Z**Woi**|**li**| **to**| twarz|?| **czy**|**li**| kobyła|?|.↵
Niech| **mi**|ę| urod|ą| t|ą| nie| omyla|/| **Tw**|**á**|**sz**| **Tw**|**á**|**sz**| **Tw**|**ász**| **to**| post|aw|ą|?| **twoi**|**sz**| **to**| lice|?|.↵
Czy| w| **dzie**|**ści**|**ę**| lat| **m**|**o**|**d**|y| **ż**|**r**|**e**|**b**|**i**|**c**|**e**|.↵
Cwoy| **ze**| **to**| pyszczek|?| **czy**|**li**| kiernożi|/| Krzyw|y| Załom|kiem| **co**| ludz|iom| grożi|.↵
Twoi|**sz**| **to**| **oczy**|?| **czy**|**li**| od| sowy|/| w| Wi|ę|zi|ę|**ni**|u| trzymasz|/| n|ą| w|n|ą|**tr**|z|u| głow|y|.↵
Brwiczki| przypr|awne|/| Przyemna| sztuk|ą|/| Ale| t|ą|k| wdzi|ę|cz|ne|/| i|**á**|k| v| Borsuk|**á**|.↵
Nos| i|**á**|ko| śiekacz|/| **á**| plec| tey| b|**á**|rwy|/| **i**|**á**|ko| M|**á**|**l**|**á**|r|z|e| m|**á**|l|u|**i**|**ą**| i|**á**|rwy|.↵
Z|**á**|**b**|**k**| z| Hebanu|/| wiedz| **kt**|ory| **k**|ę|dy|/| N|**á**|**d**|**p**|**u**|**s**|**t**|**o**|**s**|z|**á**|y| ob|**á**|d|w|**á**| r|z|ę|**d**|y|.↵
Ten| defekt| t|**á**|i|**ą**|c|/| w|**á**|rgi| **z**|**á**|**t**|y|**k**|**á**|/| Y| i|u|ż| nie| mow|/| **l**|e|ć| c|**á**|le| **k**|s|z|y|**k**|**á**|.↵
Gdy| **p**|o| **F**|**r**|**á**|**n**|**c**|**u**|**s**|**k**|u| w|**á**|rkocz| rozwini|e|/| **B**|**ę**|**d**|**z**|i|e| pod| pach|**ą**| **i**|**á**|ko| v| świnie|.↵
Kędy| też| **h**|y|**s**|**s**|o|/| sztuk|**ą**| z|**á**|bieży|.↵
A| cudz|**ą**| śier|ci|**ą**| głow|ę| n|**á**|e|ży| W| tańcu| i|**ą**| **w**|**i**|**d**|z|e|ć|/| **w**|**i**|**e**|y| r|**á**|czym| chodzie|.↵

Twa

twój adj sg:nom:f:pos

z

z qub

Twa

twój adj sg:nom:f:pos

z

sz qub

Twarz

twarz subst sg:nom:f

to

to pred

postawa

postawa subst sg:nom:f

?

? interp

Dziękuję za uwagę

