

Doświadczenia z prac nad Korpusem tekstów polskich z XVII i XVIII wieku

zasady doboru tekstów, zakres
znakowania, sposób
udostępniania



Włodzimierz Gruszczyński
IJP PAN
e-mail: wlodekiewa@poczta.onet.pl



O potrzebie korpusu...

- W Pracowni Historii Języka Polskiego XVII i XVIII wieku powstaje słownik polszczyzny tego okresu.
- W latach 1954-1990:
 - opracowanie koncepcji,
 - ustalenie kanonu źródeł,
 - ekscerpcja, w wyniku której powstało (według ówczesnych szacunków) ponad 3 mln kart z cytatami, dających podstawę do utworzenia ponad 80 tys. haseł słownikowych.



O potrzebie korpusu...

- 1991-1993 – wznowienie ekscerpcji:
 - 42 tys. nowych kart,
 - prawie 3700 nowych haseł.
- 1996 – wydanie zeszytu próbnego.
- 1999-2004 – wydanie 5 zeszytów I tomu (na literę A).
- 2004 – zmiana kierownictwa Pracowni, poszukiwania nowych metod pracy.



O potrzebie korpusu...

- W 2005 r. na Festiwalu Nauki Polskiej postawiłem pytanie:

Jak przyspieszyć i ułatwić prace nad słownikiem?

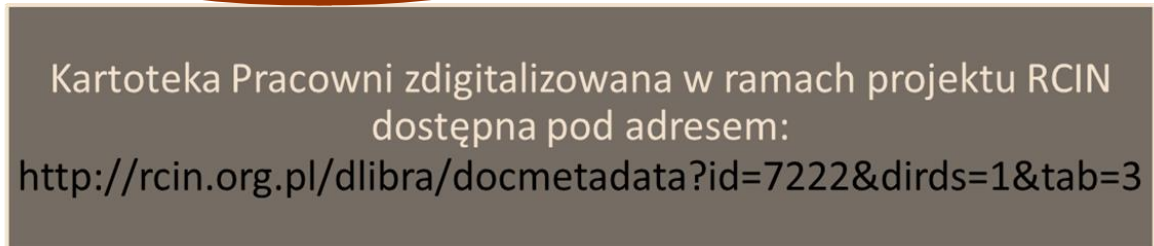
- Rozpatrywałem następujące możliwości:
 - Ułatwić dostęp do materiałów źródłowych!
 - Rozwiązać problem fiszek (Ale jak?!).
 - Przenieść jak najwięcej źródeł na nośnik elektroniczny!
 - **Stworzyć KORPUS tekstów średniopolskich.**
 - Stworzyć narzędzia ułatwiające pisanie artykułów hasłowych.



Ułatwianie dostępu do źródeł

- Współpraca z bibliotekami tworzącymi zasoby internetowe.
- Dygitalizacja słowników – przede wszystkim *Thesaurusa* Knapskiego (wydany na CD przez BUW).
- Stworzenie Cyfrowej Biblioteki Druków Ulotnych Polskich i Polski Dotyczących (obecnie dostępnej w ramach FBC).

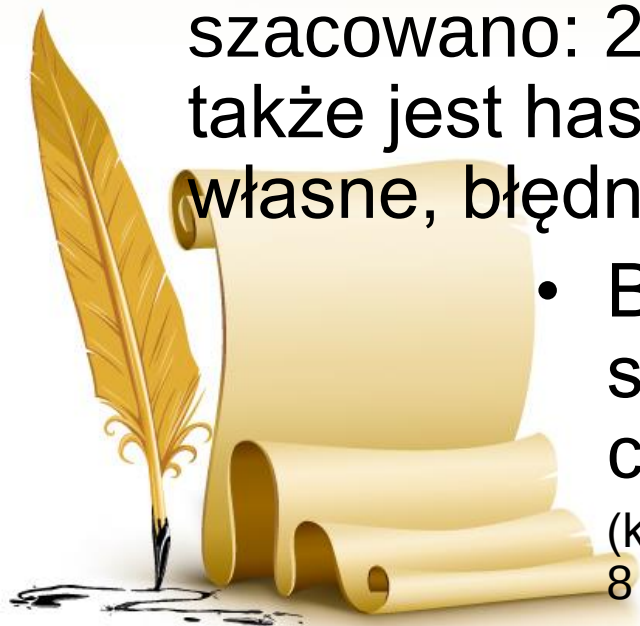




Unowocześnienie warsztatu pracy

- Wszystkie – z wyjątkiem korpusu – „nierealne” pomysły wymienione na Festiwalu Nauki zostały zrealizowane w stosunkowo krótkim czasie.
- Dygitalizacja fiszek (dokonana w ramach grantu RCIN) dowiodła, że jest ich mniej niż szacowano: 2,8 mln, a nie ponad 3 mln, mniej także jest haseł (wiele to warianty, nazwy własne, błędne zapisy).
- Baza materiałowa *S-XVII* okazała się dużo uboższa niż baza *S-XVI* czy *SJPDor*.

(kartoteka każdego z tych dwóch słowników zawiera ok. 8 mln fiszek).



O potrzebie korpusu...

- Korpus tekstów polskich XVII i XVIII wieku dałby możliwość automatycznego wygenerowania indeksów form, a może nawet leksemów.
- Korpus w pewnym stopniu zastąpiłby kartotekę, a w znacznym – uzupełniłby ją.
- Korpus przyspieszyłby prace nad słownikiem, a w pewnym stopniu go zastąpił (!).



PolDiLemma,

czyli czy korpus tekstów średniopolskich już istnieje?

Plik Edycja Widok Historia Zakładki Narzędzia Pomoc

UdS :: PolDiLemma: Licence x korba korpus - Szukaj w Go... x KORBA - Computational Li... x +

https://fedora.clarin-d.uni-saarland.de/poldilemma/license.html

CLARIN-D
GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

UNIVERSITÄT
DES
SAARLANDES

HUMBOLDT-UNIVERSITÄT
ZU BERLIN

Welcome to the Middle Polish Diachrone Lemmatized Corpus

PolDiLemma-Corpus

The PolDiLemma corpus is in the public domain, because the copyright of the original authors has expired.

The authors of the annotations release their work into the public domain. This applies worldwide. In some countries this may not be legally possible; if so: they grant anyone the right to use this work for any purpose, without any conditions, unless such conditions are required by law.

All texts (free licenses) are gathered from Federacja Bibliotek Cyfrowych (Digital Library Federation). The Middle Polish texts illustrate the history of the language and give the opportunity to explore some first-hand evidence of the development of Polish in its historical context.

PolDiLemma-Tool

PolDiLemma-Tool

W3C XHTML 1.0

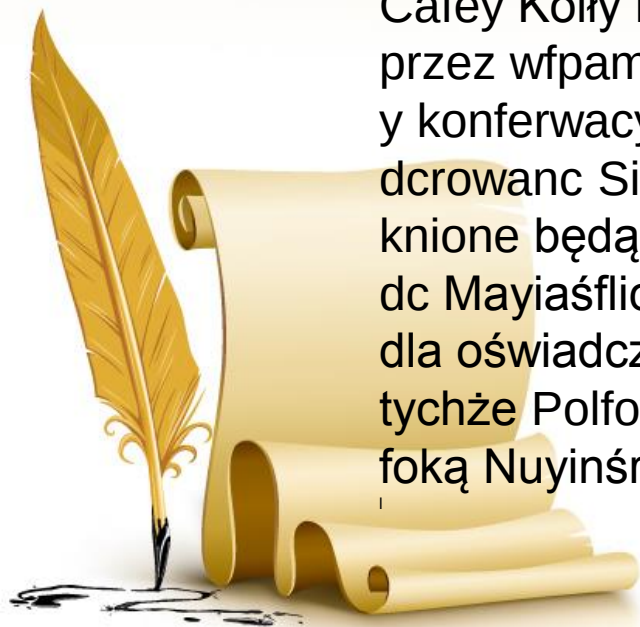
www.uni-saarland.de

Pulpit Biblioteki Włodzimierz Komputer

09:22
2016-05-12

PolDiLemma – próbka tekstu

Stany Koronne y Wielkiego Xięstwa Litewkiego
ziednoczy wf/.y Gę wę/fem Koi:(edetacyi Genc-
ralncy, dla Otrzymania Wolności Narodowcy,
iako y dla naprawy defektów &abuftitm, lub
ulepfzenia wlzcikiego, ktorcbv Gę w Oy czy/nic
naizey potrzebnym pokaziJo ; a peftępek takowy Patryety-
czny będąc wparty oc Nayi^nicvfzcy Impcratrirowy JMci
Cafey Koily i, fzczcrey A.li.inrfri /Przyjaciółki Nafzcy (bo tego
przez wfpmafe intcreflbwanie fię ku Dobni Rzeczypofpolitcy,
y konferwacyi Wolności nafzey dowody dała) więc też Skorifc-
dcrowanc Siany Oboyga Naredow, tak nieoboigtnemi przeni-
knione będąc przyjaźni y dobrego faficdzstwa dowodami, fnltrme
dc Mayiaśflicyfzcy Impcratorowy Jcy Mci wyff.fy Polclllwo,
dla oświadczenia szczerego podziękowania, y uprafzania przez
tychże Polfow o kontynuację pomocy przyiaciclikiey, y o wy-
foką Nuyinśr.icyfzey ImpcraiorOwy Gwarancyą, dla niewzru-



Starania o grant

- Pierwsza próba (w MNiSzW) zakończona rezygnacją z przyznanej dotacji (bo była żenująco mała).
- Druga próba zakończona sukcesem:
 - Projekt badawczy „Elektroniczny korpus tekstów polskich XVII i XVIII w. (do 1772 r.)”.
 - Finansowanie projektu: Narodowy Program Rozwoju Humanistyki na lata: 2013-2018.
 - Jednostka koordynująca: IJP PAN.
 - Współpraca: IPI PAN.
 - Kierownik: Włodzimierz Gruszczyński.
 - Koordynator: Renata Bronikowska.
 - Kryptonim: KORBA (**KOR**pus **BA**rokowy).



Najważniejsze założenia

- Chronologiczne rozszerzenie NKJP.
- Planowana objętość: 12M segmentów
 - » mała w porównaniu z NKJP (300M w korpusie zrównoważonym),
 - » relatywnie duża jak na korpus historyczny, por.:
 - 1,2M – szwedzkie podkorpusy z okresu Nysvenska, tzn. 1520–1850,
 - 11M – serbski korpus historyczny (KSJ),
 - 100M – Corpus del Español (lata: 1200-1900 r.),
 - 406M – Corpus of Historical American English (COHA) (lata: 1810-2000).



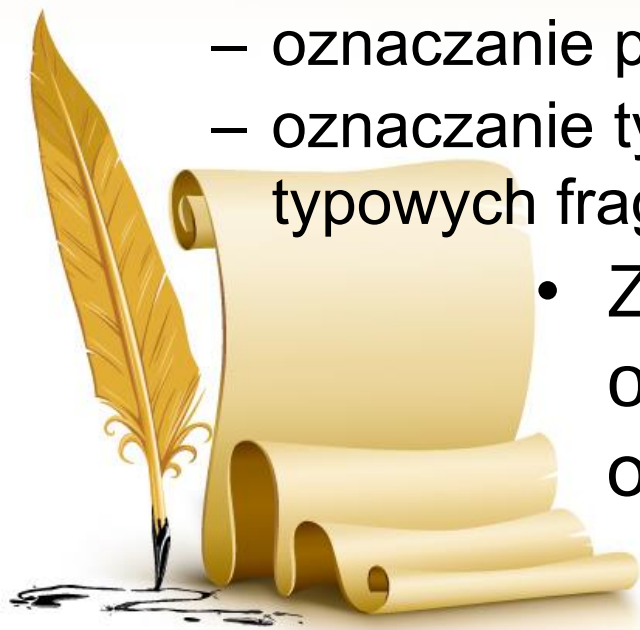
Najważniejsze założenia – metadane

- Dokładne dane bibliograficzne tekstu.
- Charakterystyka stylistyczno-genologiczna.
- Charakterystyka socjolingwistyczna i geolingwistyczna (informacje o autorze, wydawcy, tłumaczu).



Najważniejsze założenia – znakowanie

- Znakowanie strukturalne umożliwiające dokładną lokalizację cytatu w tekście:
 - paginacja (uwzględnienie różnych systemów),
 - oznaczenia granic rozdziałów, części, ksiąg itp.,
 - oznaczanie notek marginesowych, przypisów itp.,
 - oznaczanie podpisów pod ilustracjami;
 - oznaczanie typowych części, np. dedykacyj, typowych fragmentów strony tytułowej itp.
- Znakowanie językowe, tzn. oznaczenie wszystkich wtrętów obcych, z podziałem na języki.



Znaczniki strukturalne

[POCZĄTEK DOKUMENTU] - [KONIEC DOKUMENTU]

[POCZĄTEK STRONY] - [KONIEC STRONY]

[POCZĄTEK POZIOMU 1] - [KONIEC POZIOMU 1]

[POCZĄTEK STRONY TYTUŁOWEJ] - [KONIEC STRONY TYTUŁOWEJ]

[POCZĄTEK STRONY POTYTUŁOWEJ] - [KONIEC STRONY POTYTUŁOWEJ]

[POCZĄTEK MOTTA] - [KONIEC MOTTA]

[POCZĄTEK PRE-TEKSTU] - [KONIEC PRE-TEKSTU]

[POCZĄTEK POST-TEKSTU] - [KONIEC POST-TEKSTU]

[POCZĄTEK LISTY] - [KONIEC LISTY]

[POCZĄTEK INDEKSU] - [KONIEC INDEKSU]

[POCZĄTEK SPISU TREŚCI] - [KONIEC SPISU TREŚCI]

[POCZĄTEK NOTKI] - [KONIEC NOTKI]

[ILUSTRACJA]

[WZÓR MATEMATYCZNY]

[ZAPIS NUTOWY]

[INNA PRZERWA W TEKŚCIE]

[FRAGMENT NIECZYTELNY]

[TEKST W JĘZYKU OBCYM]

[ALFABET NIEŁACIŃSKI]



Znaczniki tekstowe

Tytuł publikacji	Podtytuł części
Powtórzony tytuł	<i>Streszczenie części</i>
Podtytuł publikacji	Podpis pod rysunkiem
<i>Autor</i>	Oznaczenie składki
Miejsce wydania	<i>Numer strony</i>
Drukarnia	<i>Kustosz</i>
<i>Rok wydania</i>	<i>Żywa pagina</i>
Inny element strony tytułowej	<i>Numer rozdziału</i>
<i>Tłumacz</i>	<i>Numer wersetu</i>
<i>Język oryginału</i>	<i>Element dodany w Biblii Gdańskiej</i>
Numer wydania	Tekst odredakcyjny
Tytuł części	Tekst współczesny



Oznaczenia języków obcych

Arabski	Niemiecki
Czeski	Południowosłowiański
Francuski	Wschodniosłowiański
Grecki	Skandynawski
Hebrajski	Turecko-tatarski
Hiszpański	Węgierski
Litewski	Włoski
Lacina	Inny



Najważniejsze założenia

– znakowanie morfosyntaktyczne

- Znakowaniem morfosyntaktycznym wprowadzanym ręcznie objęty będzie podkorpus o długości 0,5M segmentów. Całość będzie oznakowana automatycznie.
- Zestaw znaczników (tagset) budowany jest na podstawie zestawu zastosowanego przy znakowaniu NKJP.
- Konieczne jest wprowadzenie zmian umożliwiających oznakowanie kategorii i/lub wartości niewystępujących we współczesnej polszczyźnie.
 - Wskazane jest wprowadzenie znaczników zgodnych z tradycją opisu historycznego i umożliwienie wyszukiwania według tradycyjnych kategorii.



Pozyskiwanie tekstów

- Rękopisy, starodruki z epoki:
 - najbardziej wiarygodne,
 - zwykle niedostępne w tekstowej formie elektronicznej,
 - drogie w pozyskiwaniu.
- Wydania z XIX, XX, XXI w.:
 - mało wiarygodne (ortografia, gramatyka),
 - czasem dostępne w formie elektronicznej,
 - wielu dotyczy ograniczenie z powodu prawa autorskiego.
- Edycje internetowe (mało wiarygodne)
 - pozytywny przykład – *Wolne lektury*.

przepisywane przez „skryptorów” w formacie MS Word ze specjalnymi stylami i skrótami klawiaturowymi



Przykład tekstu przepisanego i oznakowanego strukturalnie oraz językowo

[POCZĄTEK·DOKUMENTU]¶

[POCZĄTEK·STRONY·TYTUŁOWEJ]¶

PIECHOTNE·CWICZENIE¶

Albo· WOIENNOSC· PIESZA· Ktorą· Láćinnicy·
Pedestrem· Militiam· názywáią· Wodzom·
Pułkownikom· y· wszelkíey· Woíenney·
Stárszynie· Lubo· ktoszkolwíek· íest· Woíenney·
spráwy· Míłóśnikamiem·¶

Do·wíadomości·podána·¶

Przez·BLAZEIA·LIPOWSKIEGO¶

Nakładem·lerzego·Forstera·l.K.M·Bibli·opoli·<?>¶

W·KRAKOWIE¶

W·Drukárni·v·Wdowy·Lukaszá·Kupiszá·l.K.M·¶

Typogr·Roku·Panskíego·1660·¶

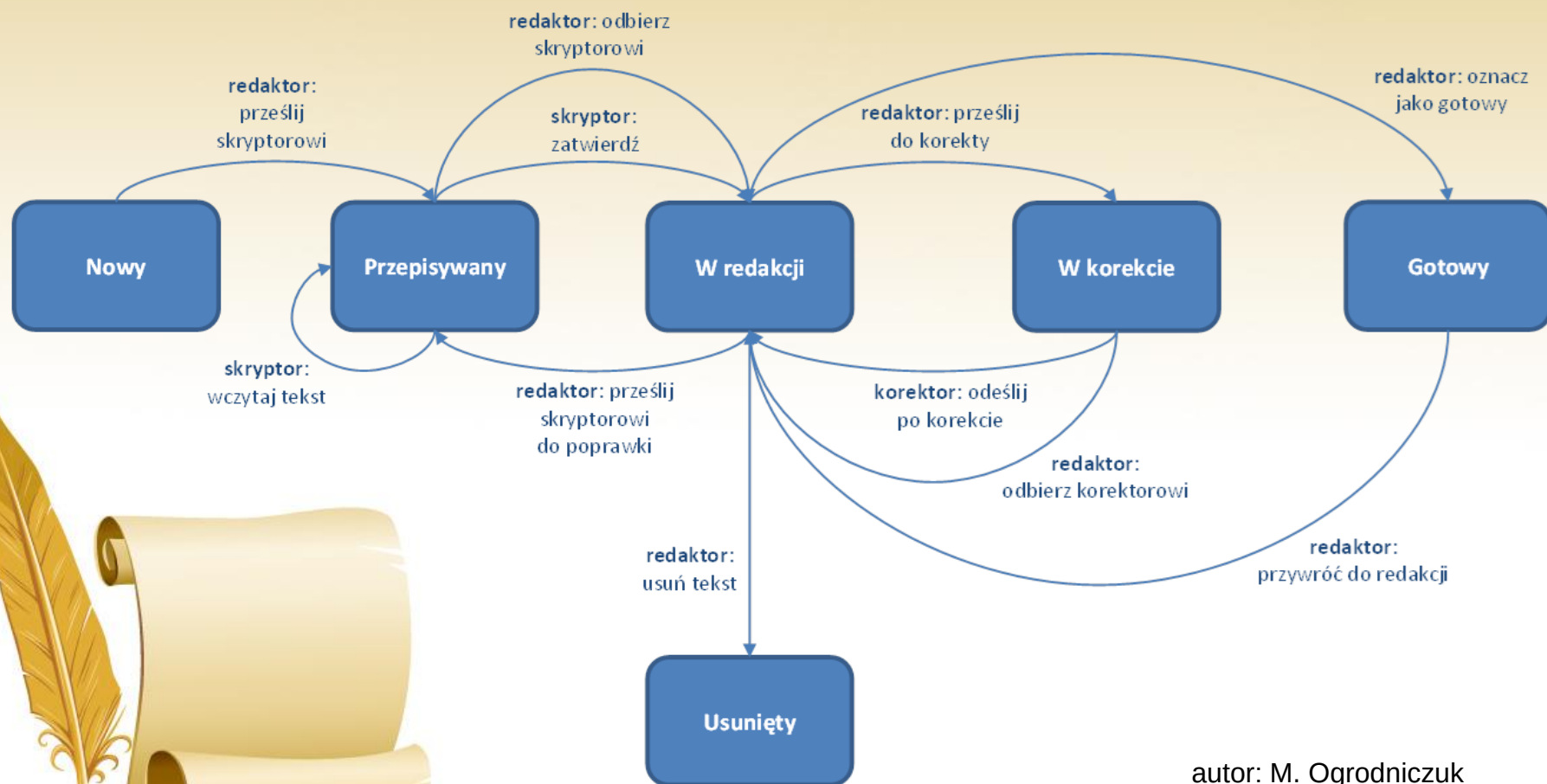
[KONIEC·STRONY·TYTUŁOWEJ]¶



Przykład tekstu w formacie XML

```
- <teiCorpus>
  <xi:include href="KORBA_header.xml"/>
- <TEI>
  <xi:include href="header.xml"/>
- <text>
  - <front>
    <pb/>
  - <titlePage xml:id="txt_1-titlePage">
    - <docTitle xml:id="txt_1.1-docTitle">
      <titlePart type="main" xml:id="txt_1.1.1-titlePart">PIECHOTNE CWICZENIE</titlePart>
      - <titlePart type="sub" xml:id="txt_1.1.2-titlePart">
        Albo WOIENNOSC PIESZA Ktorą Lácinnicy
        <foreign xml:lang="la" xml:id="txt_1.1.2.1-foreign">Pedestrem Militiam</foreign>
        nazywáią. Wodzom, Pułkownikom, y wszelkiey Woienney Stárszynie. Lubo ktoszkolwiek iest Woienney sprawy Miłośnikiem.
      </titlePart>
    </docTitle>
    <note xml:id="txt_1.2-note">Do wiadomości podána.</note>
    <docAuthor xml:id="txt_1.3-docAuthor">Przez BLAZELA LIPOWSKIEGO</docAuthor>
  - <note xml:id="txt_1.4-note">
    Nakładem Ierzego Forstera, I.K.M. Bibli
    <uncertain reason="unreadable" xml:id="txt_1.4.1-uncertain">opoli.</uncertain>
  </note>
  - <docImprint xml:id="txt_1.5-docImprint">
    <pubPlace xml:id="txt_1.5.1-pubPlace">W KRAKOWIE</pubPlace>
    <publisher xml:id="txt_1.5.2-publisher">W Drukárni v Wdowy Łukaszá Kupiszá, I.K.M. Typogr.</publisher>
    <docDate xml:id="txt_1.5.3-docDate"> Roku Panskiego 1660.</docDate>
  </docImprint>
</titlePage>
```

Etapy pracy przy transkrypcji tekstu za pomocą Wczytywacza



autor: M. Ogrodniczuk

Wczytywacz tekstów

- Ułatwia przekazywanie tekstów między redaktorem a skryptorem (korektorem).
- Umożliwia redaktorowi tworzenie metryczki tekstu, czyli wprowadzanie bardzo szczegółowych metadanych.
- Wychwytuje błędy formalne w znakowaniu,
 - np. sygnalizuje brak znacznika [koniec strony], jeśli nie ma go między dwoma znacznikami [początek strony].
- Konwertuje tekst na format XML (zgodny z TEI).
 - Sporządza raporty dotyczące liczby tekstów i słów należących do danej kategorii.
 - Wyszukuje zadany ciąg liter we wczytanych tekstach.



Wczytywacz – status tekstów

Plik Edycja Widok Historia Zakładki Narzędzia Pomoc

UdS :: PolDiLemma: Home x korba korpus - Szukaj w Go... x KORBA - Computational Li... x KORBA — wczytywacz teks... x +

korba.nlp.ipipan.waw.pl

KORBA — wczytywacz tekstów

włodek (Redaktor)

Strona główna
Nowy tekst
Wyszukiwarka
Raporty

Pracuj jako:
skryptor
korektor

Zgłoś błąd
Zmień hasło
Wyloguj

Wczytanych tekstów: **672**,
zawierających w sumie
10573778 (10439083 wg TEI) słów.

Zakończonych tekstów: **515**,
zawierających w sumie
6598644 (6516810 wg TEI) słów.

Identyfikator	Metadane	Stan	Redaktor	Skryptor	Korektor	Słów w doc (w TEI)	Znaków	Tekst	Akcje
PetrSInst	Sebastian Petrycy <i>Instrukcja albo nauka jak się sprawować czasu moru</i> Mikołaj Lob Kraków 1613	Gotowy	renata			11074 (10941)	72515 (40,2str)	nagłówek źródło podgląd XML	!
BirkEgz	Fabian Birkowski O Egzorbítancjach kazania dwoje Andrzej Piotrkowczyk Kraków 1632	Gotowy	renata	renata		8657 (8482)	57010 (31,6str)	nagłówek źródło podgląd XML	!
RysProv	Salomon Rysiński <i>Proverbium polonicorum</i>	Gotowy	renata	renata		16616 (16359)	103500 (57,5str)	nagłówek źródło podgląd XML	!

Pulpit Biblioteki Włodzimierz Komputer

PL 12 10:41 2016-05-12

Wczytywacz – metryczka tekstu

Utwórz nowy tekst

Id: *

Tytuł: *

Autor: * ☐ Anonimowy

Tłumacz: ☐ Anonimowy

Drukarnia:

Rok:

Rok wydania niepewny
(wydano nie wcześniej, niż podany rok): ☐

Typ mowy: *

Rodzaj: *

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

Tematyka:

☐ alchemia ☐ anatomia ☐ architektura ☐ astrologia ☐ astronomia ☐ biologia ☐ botanika
☐ budownictwo ☐ chemia ☐ egzotyka ☐ ekonomia ☐ filozofia ☐ fizyka ☐ geografia ☐ gospodarstwo
☐ gramatyka ☐ historia ☐ języki ☐ kulinaria ☐ matematyka ☐ medycyna ☐ mitologia ☐ miłość
☐ muzyka ☐ myślistwo ☐ obyczajowość ☐ polityka ☐ prawo ☐ religia ☐ retoryka ☐ wojskowość
☐ zielarstwo ☐ zoologia ☐ żeglarsstwo

Poetyka żartu: ☐

Wczytywacz – raport

typ mowy:

rodzaj:

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

tematyka:

- ☐ alchemia ☐ anatomia ☐ architektura ☐ astrologia ☐ astronomia ☐ biologia ☐ botanika ☐ budownictwo ☐ chemia
- ☐ egzotyka ☐ ekonomia ☐ filozofia ☐ fizyka ☐ geografia ☐ gospodarstwo ☐ gramatyka ☐ historia ☐ języki
- ☐ kulinaria ☐ matematyka ☐ medycyna ☐ mitologia ☐ miłość ☐ muzyka ☐ myślistwo ☐ obyczajowość ☐ polityka
- ☐ prawo ☐ religia ☐ retoryka ☐ wojskowość ☐ ziołarstwo ☐ zoologia ☐ żeglarstwo

poetyka
żartu: ☐

region:

50-lecie:

[Generuj raport](#)

Znaleziono:

Tekstów: 5

Segmentów: 90891

Wczytywacz – tymczasowa wyszukiwarka

[Powrót do strony głównej](#)

Wyszukiwarka

Hasło: *

Wyszukaj

1

2

3

4

Link

[BotłeczRel IV 84244 korbaczow:](#)

[BujnDroga 40801 korbacz/](#)

[BystrzInfElem 6926 szkorbutu](#)

[BystrzInfRóżn 2263 korba,](#)

[BystrzInfRóżn 2276 korba:](#)

[BystrzInfZup 4499 szkorbutow,](#)

[ChmielAteny III 69268 korbé.](#)

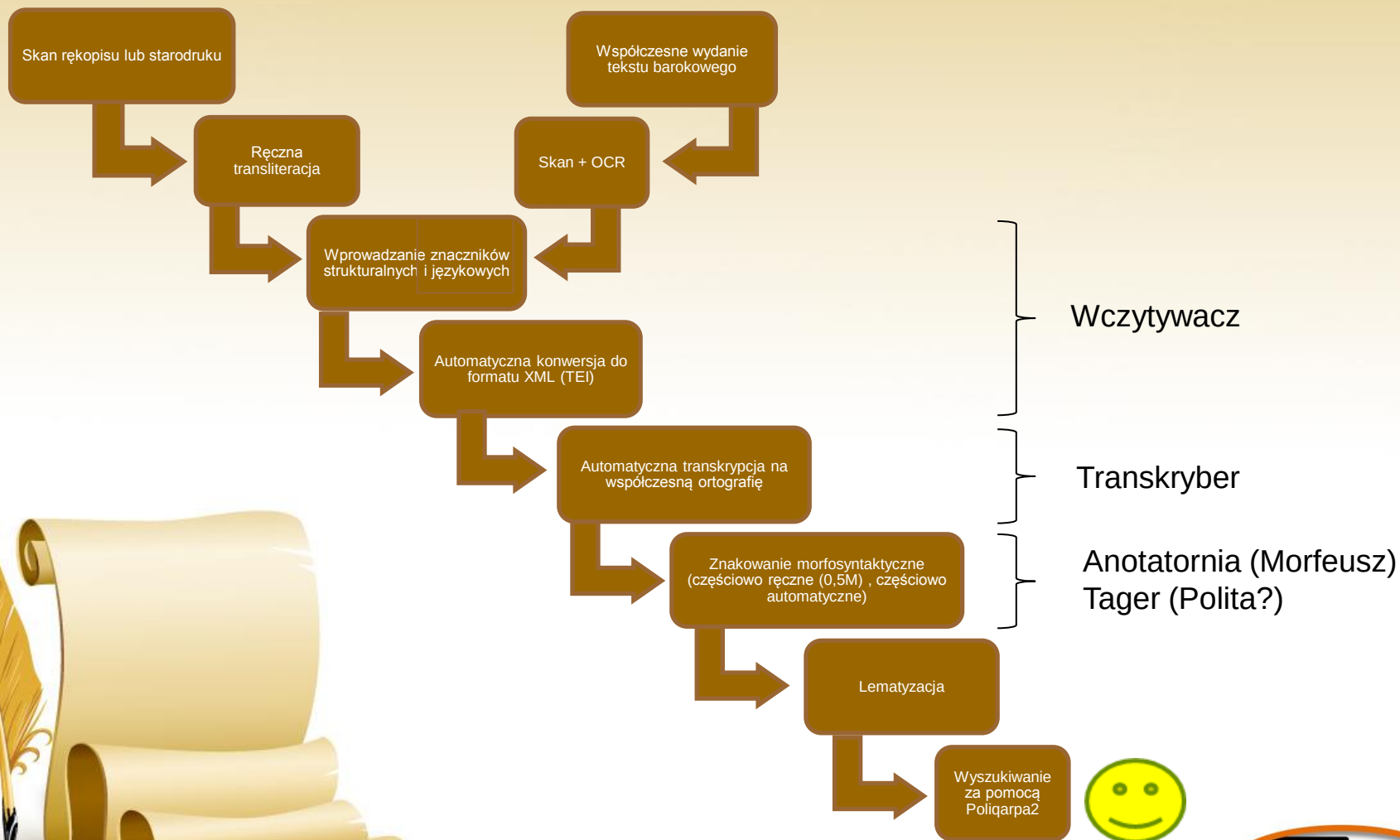
[ChmielAteny III 80797 korbacz.](#)

[CiachPrzyp 7552 skorbućie](#)

[DružZbiór 50587 korbáčzem](#)



Schemat tworzenia korpusu



Problem zrównoważenia i reprezentatywności korpusu

- Trudności zarówno teoretyczne, jak i praktyczne większe niż w wypadku korpusu tekstów współczesnych, zwłaszcza wielkich.
- Ograniczona wiedza o strukturze zbioru tekstów funkcjonujących w Polsce w XVII-XVIII w. oraz o czytelnictwie w epoce.
- Brak niektórych typów tekstów (w szczególności – mówionych).
- Problem ilościowy: włączenie wielkich tekstów w całości do relatywnie małego korpusu doprowadzi do braku zrównoważenia, por. dwa poniższe teksty to potencjalnie 10% planowanego korpusu:
 - *Biblia Gdańska* – ok. 742 000 segmentów
 - *Zielnik Syreniusza* – ok. 486 000 segmentówok. 1,2 mln segmentów
- Problem dostępności tekstów (wiele popularnych w epoce nie dochowało się), możliwości ich pozyskania (trudność w znalezieniu osób umiejących przepisywać rękopisy).

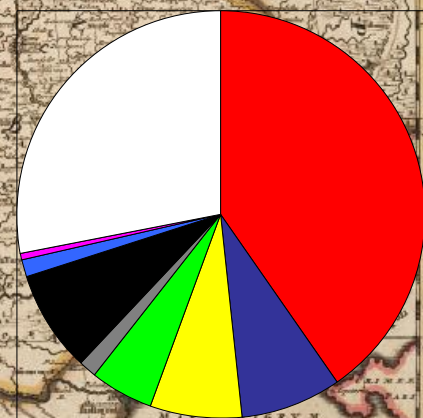
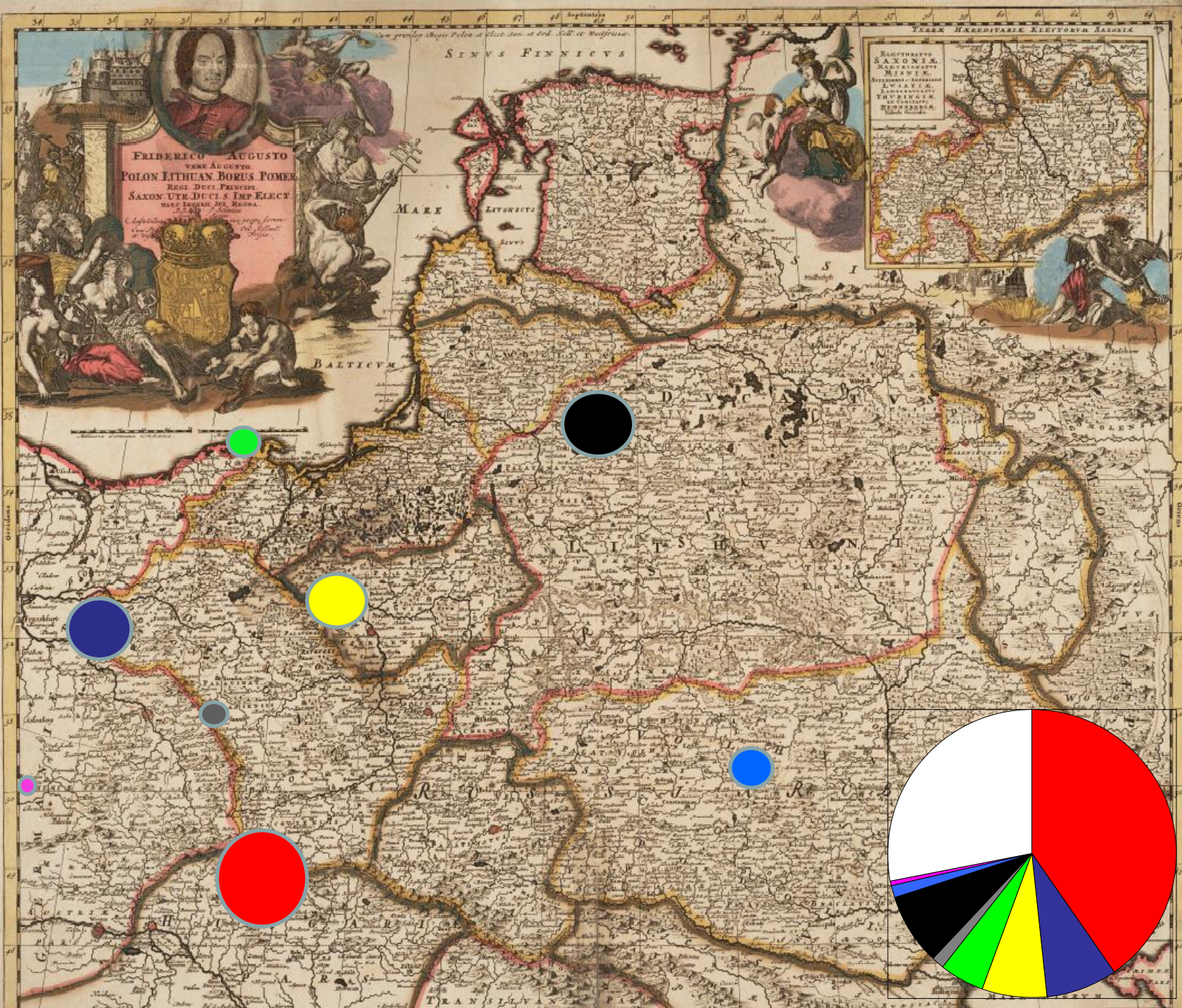


Reprezentatywność – cele

- W korpusie historycznym ważne jest zgromadzenie tekstów maksymalnie zróżnicowanych pod względem stylistycznym, tematycznym, chronologicznym i regionalnym.
- Dążenie do zrównoważenia i zachowania reprezentatywności korpusu może utrudniać osiągnięcie tego celu.
- Por.: D. Adamiec, *Kryteria doboru tekstów do Elektronicznego korpusu tekstów polskich z XVII i XVIII w. (do 1772 r.)*, „Prace Filologiczne” LXVII (2015), 11-20.

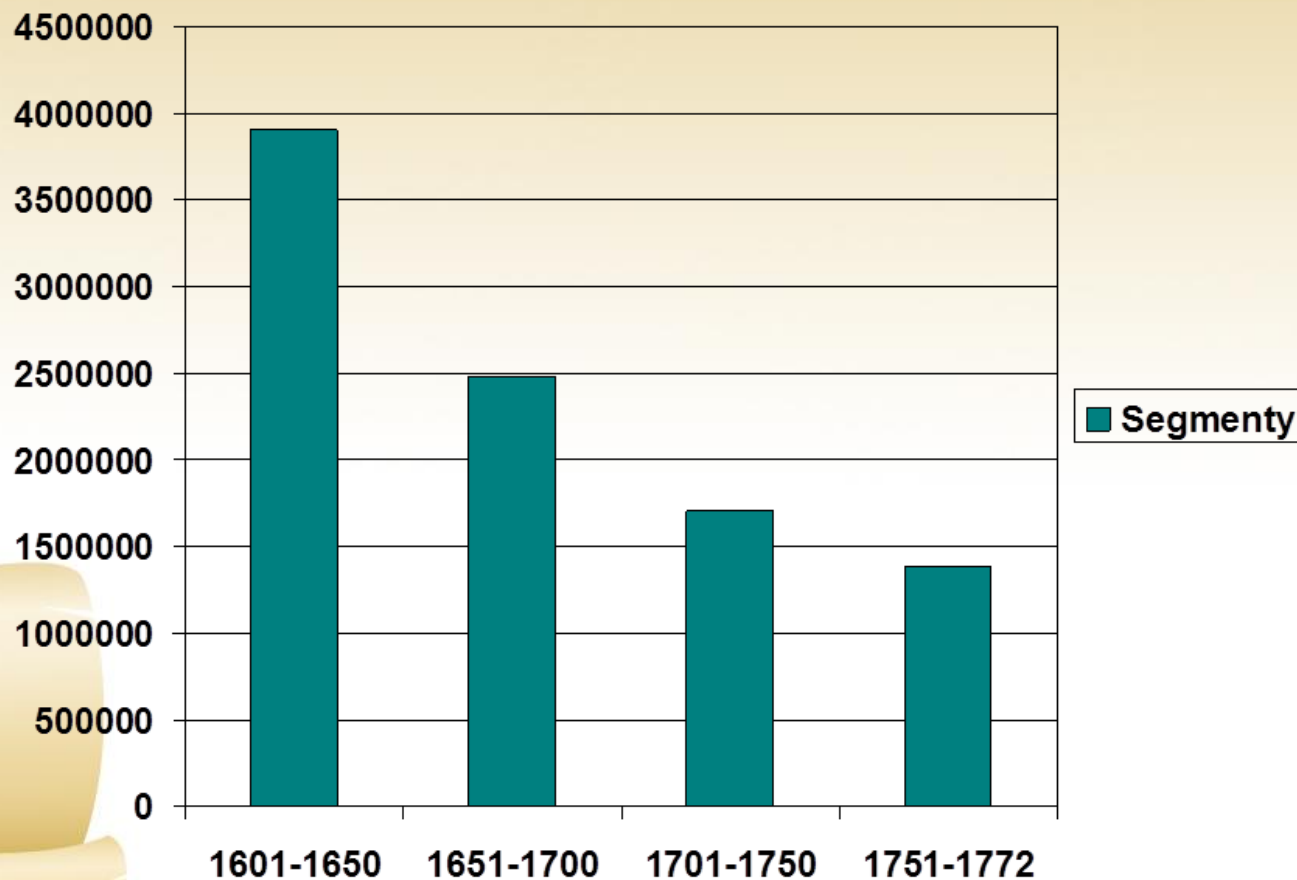


Reprezentatywność: regiony



- Małopolska
- Wielkopolska
- Mazowsze
- Pomorze i Prusy
- Śląsk
- Ziemie W. Ks. Lit.
- Ziemie Ruskie
- zagranica
- nieznanne

Reprezentatywność: chronologia



Znakowanie morfosyntaktyczne

- Przypisanie każdemu segmentowi informacji gramatycznej:
 - o jego przynależności do określonej klasy gramatycznej,
 - o wartościach kategorii gramatycznych.
- Umożliwia zaawansowane przeszukiwanie korpusu, np. wyszukanie wszystkich form fleksyjnych leksemu.
- Znakowanie Korby będzie się odbywać na wzór znakowania NKJP.
- Tagset (zbiór znaczników) będzie oparty na tagsecie NKJP, który zostanie dostosowany do potrzeb opisu polszczyzny barokowej.
- Znakowanie będzie się odbywać na tekstach transkrybowanych, a nie transliterowanych.



Konwersja do wersji transkrybowanej

transliteracja	transkrypcja
B. Nie z ^á da mi żaden te ^y rzeczy, kt ^o raby b^árzie^y owemu niżeli mnie należeć nie mi ^á ł ^á . A t ^á k śmierć ^y Kupidyn poz ^á rszy się i ^á ko bestie , posnęli w Koś ^ć iele Bacchusowym, gd ^ź ie niewyszumiawszy z przepi ^ć ia śmierć Cupidin ^o w, ^á Cupido śmier ^ć i należący przypasawszy s ^á y ^d ak do boku szły, swych należących odpr ^á wow ^á ć powinnoś ^ć i.	B. Nie za ^a da mi żaden te ^j rzeczy, kt ^ó raby barzye^j owemu niżeli mnie należeć nie mi ^a ł ^a . A ta ^k śmierć ⁱ Kupidyn poz ^a rszy się ja ^k o beztie , posnęli w Koś ^c iele Bacchusowym, gd ^z ie niewyszumiawszy z przepi ^c ia śmierć Cupidyn ^ó w, ^a Cupido śmier ^c i należący przypasawszy sa ^j da ^k do boku szły, swych należących odpr ^a wowa ^a ć powinnoś ^c i.

Znaczniki gramatyczne

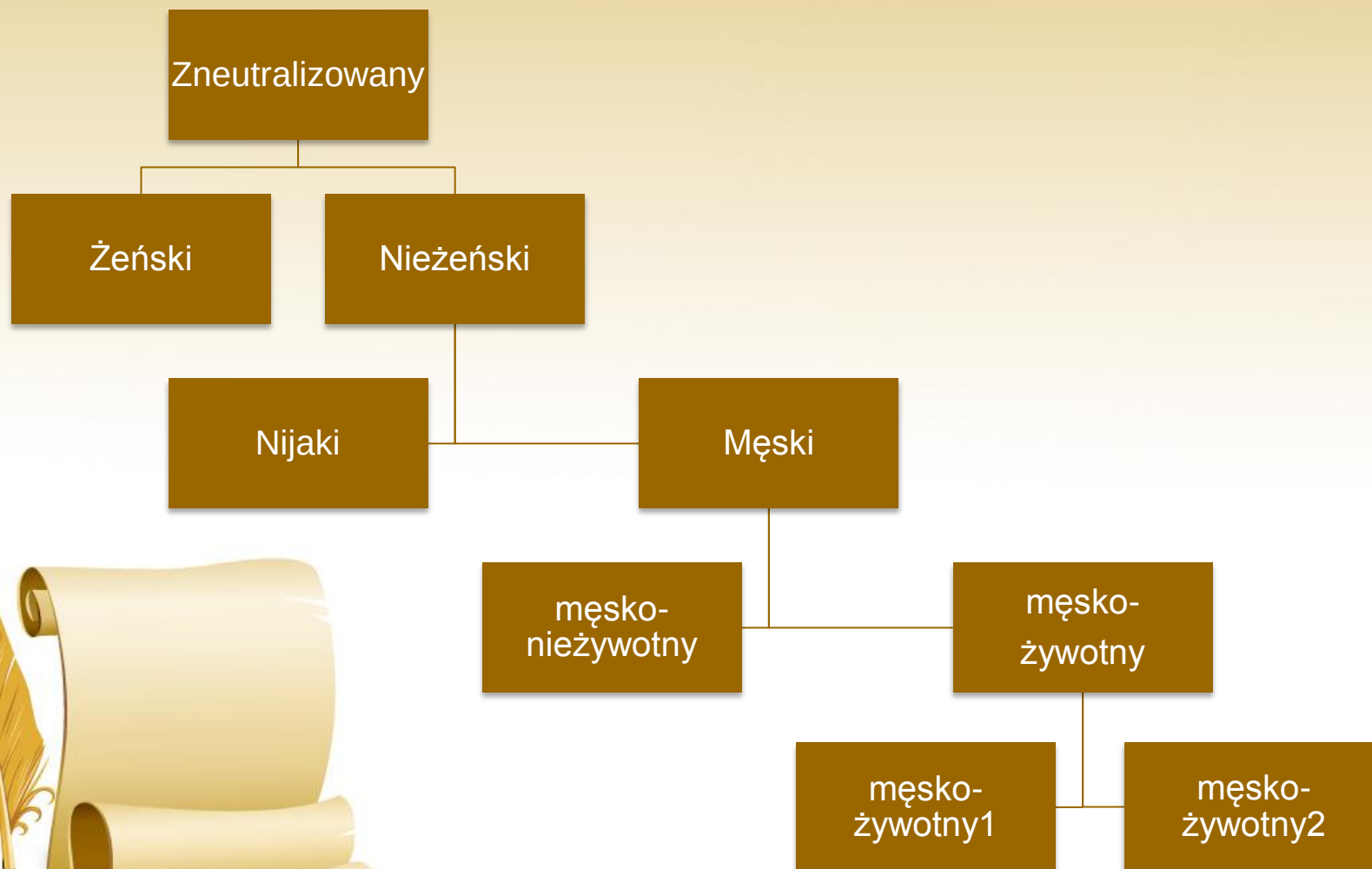
Liczba: (3 wartości)		
pojedyncza	sg	niewiasta
podwójna	du	niewieście
mnoga	pl	niewiasty
Przypadek: (7 wartości)		
mianownik	nom	woda
dopełniacz	gen	wody
celownik	dat	wodzie
biernik	acc	wodę
narzędnik	inst	wodą
miejsownik	loc	wodzie
wołacz	voc	wodo
Rodzaj: (5 wartości)		
męski osobowy (?)	m1	papież, kto, wujostwo
męski zwierzęcy	m2	baranek, walc, babsztyl
męski rzeczowy	m3	stół
żeński	f	stuła
nijaki	n	dziecko, okno, co, skrzypce, spodnie
przymnogi osobowy	p1	wujostwo
przymnogi nieosobowy	p2	skrzypce, spodnie

dodane

CO Z
tym?

dodane

Propozycja znakowania rodzaju



Obecny stan korpusu

- Wprowadzono ponad 672 teksty.
- Łączna długość tekstów ponad 10,5M segmentów, co daje ponad 12M w rozumieniu NKJP.
- Wszystkie teksty mają dokładne metryczki.
- Wszystkie teksty są oznakowane strukturalnie i językowo.
- Większość tekstów wymaga korekty, którą obecnie prowadzimy.
- Transkryber działa, uzupełniony został zestaw reguł swoistych dla naszego korpusu.
- Anotatornia jest w fazie testów.
- Morfeusz'17 jest w trakcie tworzenia.
- Udostępnienie wersji testowej – koniec 2016 r.

