Tworzenie korpusu tekstów dawnych a korpusu tekstów współczesnych: różnice teoretyczne i warsztatowe

(na przykładzie Korpusu tekstów polskich XVII-XVIII wieku)

W. Gruszczyński – R. Bronikowska

IJP PAN

Porównywane korpusy

współczesny

- *NKJP,* czyli *Narodowy Korpus Języka Polskiego* – ogólnie znany, niewymagający prezentacji, obecnie nierozbudowywany(?).

historyczny

- *KORBA*, czyli *Korpus tekstów polskich XVII i XVIII wieku (do 1772 r.)* – jeszcze nieznany, wymaga krótkiej prezentacji, jeszcze niedostępny publicznie, intensywnie rozbudowywany.

KORBA – podstawowe dane

- Projekt badawczy „Elektroniczny korpus tekstów polskich XVII i XVIII w. (do 1772 r.)”
- Kryptonim: KORBA (KORpus BArokowy)
- Finansowanie projektu: Narodowy Program Rozwoju Humanistyki na lata: 2013-2018
- Jednostka koordynująca: IJP PAN
- Współpraca: IPI PAN
- Kierownik: W. Gruszczyński
- Koordynator: R. Bronikowska

Wielkość korpusu

Korpus współczesny (NKJP)

- Pełny korpus – 1800M segmentów, w tym:
  - podkorpus zrównoważony: 300M segmentów;
  - ręcznie znakowany podkorpus (v2): 1,2M segmentów;
  - korpus *Słownika frekwencyjnego*: 0,5M.

Korpus historyczny (KORBA)

- Planowana objętość: minimum 12M segmentów
- Planowana objętość ręcznie znakowanego podkorpusu: 0,5M segmentów.
- Planowana rozbudowa w ramach grantu DARIAH.

Wielkość podobnych korpusów zagranicznych

Korpusy współczesne

- Czeski KN SYN: >1300M:
  - podkorpus zrównoważony: 100M słów;
- Słowacki KN: >829M
- Chorwacki KN: >101M (?)
- Korpus Współczesnego Języka Serbskiego: 112M.
- Bułgarski KN: 1200M
- Rosyjski (z podkorpusem historycznym): 149M
- Brytyjski KN (BNC): 100M
- Niemiecki KN: 1900M

Korpusy historyczne

- Serbski (KSJ): 11M
- Hiszpański (Corpus del Español): 100M słów (zasięg czasowy: 1200-1900 r.)
- Szwedzki: 1,2M (1520-1850 r.)
- Corpus of Historical American English (COHA): 406M (zasięg czasowy 1810-2000)

Synchronia vs. diachronia

Korpus współczesny (NKJP)

- teksty z relatywnie krótkiego okresu:
  - po 1990: 80% tekstów,
  - 1945-90: 15% tekstów,
  - przed 1945: 5% tekstów;
- z założenia nadreprezentacja tekstów współczesnych;
- teksty zestandaryzowane pod względem ortograficznym i gramatycznym (dawniejsze według współczesnych wydań).

Korpus historyczny (KORBA)

- teksty z relatywnie długiego okresu, którego poszczególne części powinny być reprezentowane równomiernie (proporcjonalnie):
  - 1601-1650,
  - 1651-1700,
  - 1701-1750,
  - 1751-1772.
- teksty ze wszystkich podokresów w zasadzie w ortografii oryginalnej.

Synchronia vs. diachronia

Korpus współczesny (NKJP)

- Brak znaczników i narzędzi umożliwiających śledzenie zmian językowych, zwłaszcza gramatycznych;
- Brak zrównoważenia chronologicznego uniemożliwia dokonywanie porównań frekwencji tych samych jednostek w różnych okresach.

Korpus historyczny (KORBA)

- Dzięki dokładnym metadanym tekstów i ich zrównoważeniu chronologicznemu możliwe będzie śledzenie zmian ilościowych różnych form gramatycznych i leksemów w poszczególnych podokresach.

Cele korpusów

NKJP

- Badania prowadzone wszystkimi metodami językoznawstwa korpusowego:
  - *corpus illustrated* (1),
  - *corpus based* (2),
  - *corpus driven* (3).
- W zakresie gramatyki przewaga badań metodą 1. i 2.
- Badania leksykalne jako podstawa leksykografii metodą 2., może też 3. (WSJP?)

KORBA

- Docelowo powinny dominować badania metodą *corpus driven*, ponieważ brak innych źródeł wiedzy o badanym systemie językowym, w szczególności brak możliwości introspekcji.
- W leksykografii (*e-SXVII*) zdecydowanie metoda *corpus driven*.

Użytkownicy (adresaci) korpusów

NKJP

- Językoznawstwo polonistyczne:
  - leksykografia (głównie *WSJP*)
  - gramatyka opisowa
  - kultura języka!
  - stylistyka/pragmatyka
  - językoznawstwo statystyczne .
- Inżynieria językowa
  - dane wykorzystywane często w innych projektach

KORBA

- Językoznawstwo polonistyczne:
  - leksykografia (*e-SXVII*)
  - gramatyka historyczna
  - onomastyka historyczna
  - chronologia (datowanie) słów.
- Edytorstwo dawnych tekstów.
- Literaturoznawstwo.

Typy tekstów i ich dostępność

NKJP

- Relatywnie łatwy dostęp do wszelkiego typu tekstów współczesnych.
- Jedyny, ale wielki problem to prawa autorskie!

KORBA

- 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).
- Brak tekstów mówionych.

Sposoby pozyskiwania tekstów

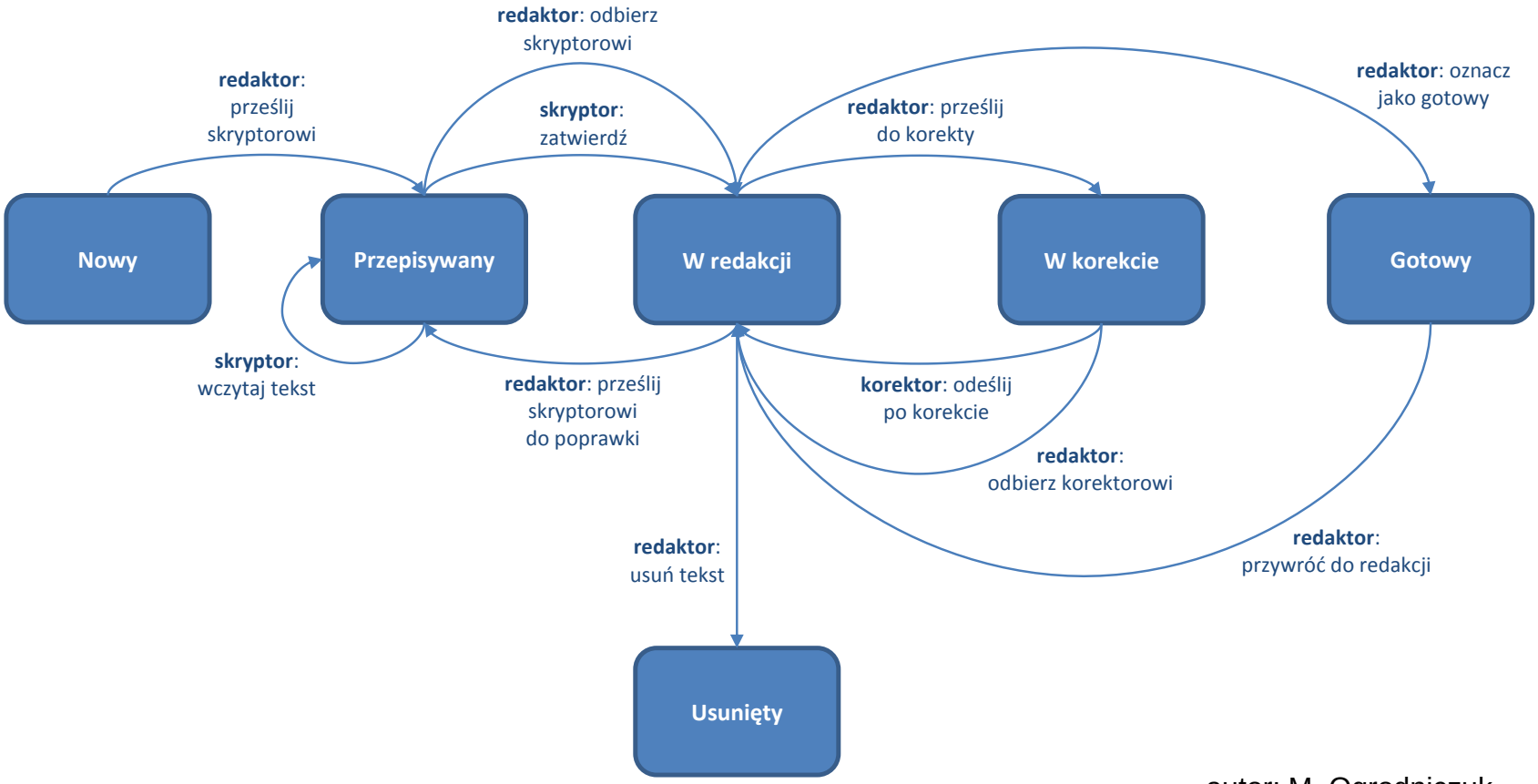
NKJP

- Pozyskiwanie wersji elektronicznej książek i czasopism;
- skanowanie i OCR tekstów pisanych niedostępnych w wersji elektronicznej;
- nagrywanie tekstów mówionych i ich transkrypcja.

KORBA

- transliteracja (przepisywanie) starodruków i rękopisów;
- pozyskiwanie wersji elektronicznej lub skanowanie i OCR tekstów dawnych wydanych współcześnie;
- „postarzanie” tekstów wydanych współcześnie (doprowadzanie ich do postaci wiernie odwzorowującej oryginał).

Etapy pracy przy transkrypcji tekstu

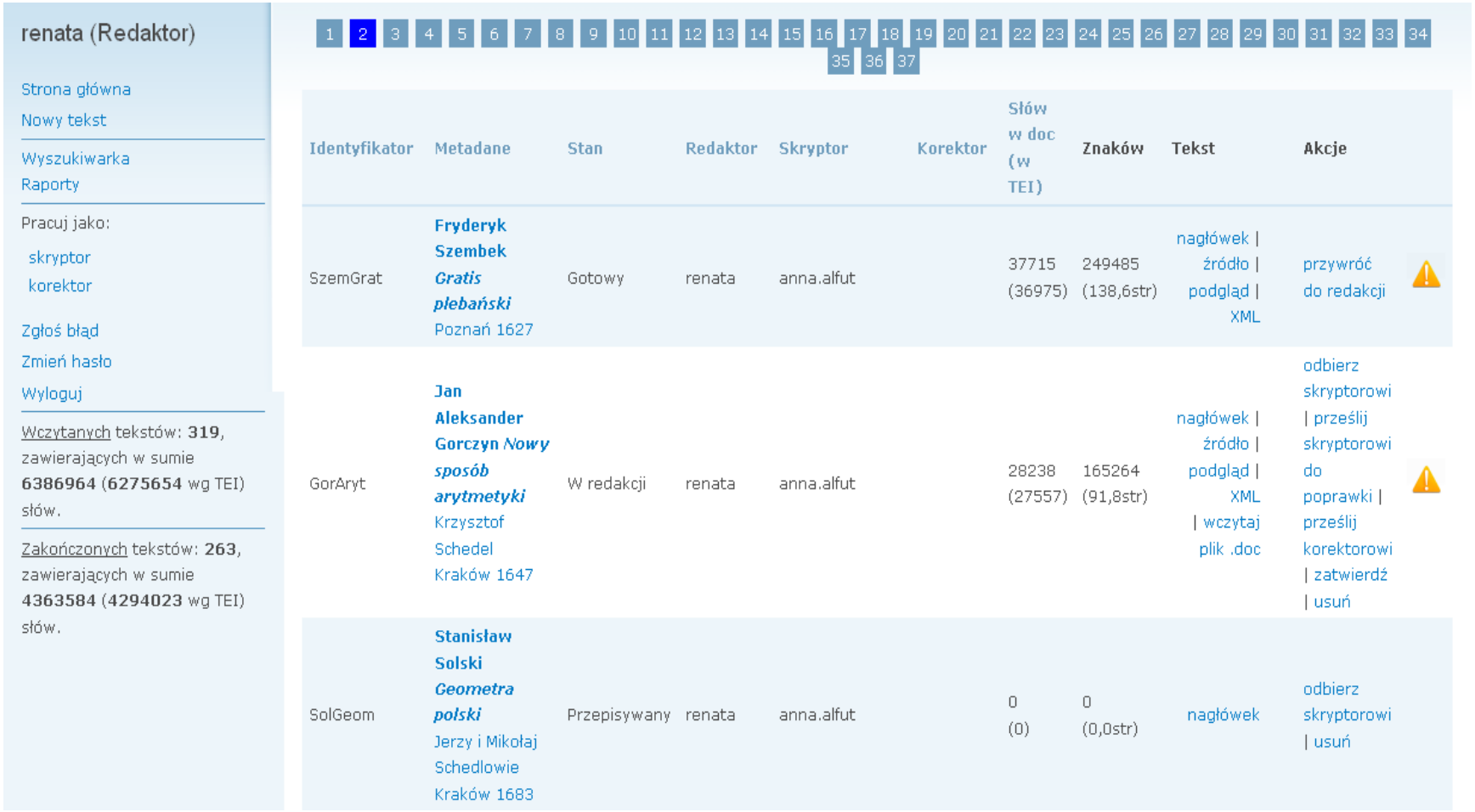


autor: M. Ogrodniczuk

Wczytywacz tekstów

- ułatwia przekazywanie tekstów między redaktorem a skryptorem;
- wychwytuje błędy w znakowaniu;
- 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.

Informacje o statusie tekstów



renata (Redaktor)

Strona główna
Nowy tekst

Wyszukiwarka
Raporty

Pracuj jako:
skryptor
korektor

Zgłoś błąd
Zmień hasło
Wyloguj

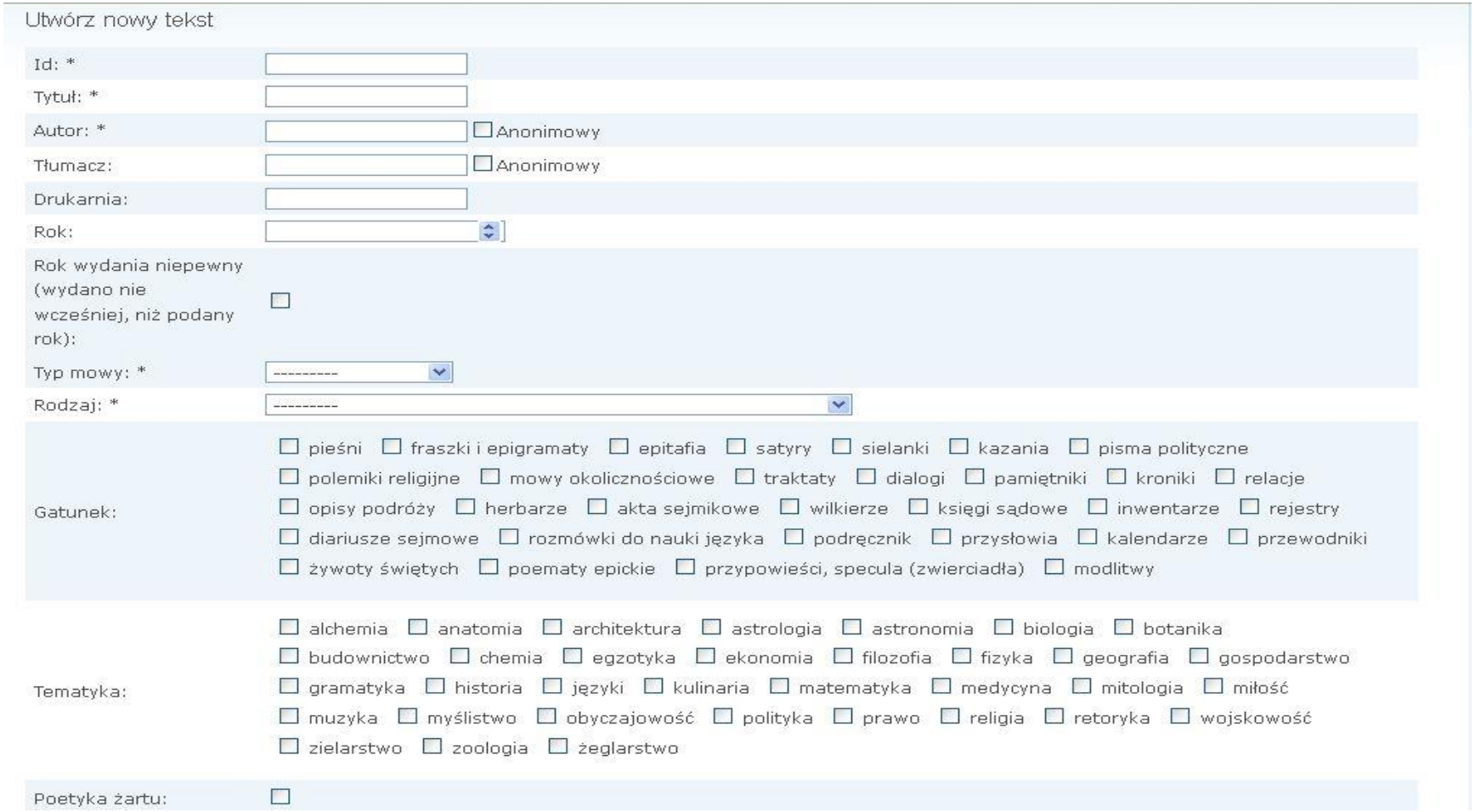
Wczytanych tekstów: **319**, zawierających w sumie **6386964** (**6275654** wg TEI) słów.

Zakończonych tekstów: **263**, zawierających w sumie **4363584** (**4294023** wg TEI) słów.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37

Identyfikator	Metadane	Stan	Redaktor	Skryptor	Korektor	Słów w doc (w TEI)	Znaków	Tekst	Akcje
SzemGrat	**Fryderyk Szembek** ***Gratis plebański*** Poznań 1627	Gotowy	renata	anna.alfut		37715 (36975)	249485 (138,6str)	nagłówek | źródło | podgląd | XML	przywróć do redakcji
GorAryt	**Jan Aleksander Gorczyn** ***Nowy sposób arytmetyki*** Krzysztof Schedel Kraków 1647	W redakcji	renata	anna.alfut		28238 (27557)	165264 (91,8str)	nagłówek | źródło | podgląd | XML | wczytaj plik .doc	odbierz skryptorowi | prześlij skryptorowi do poprawki | prześlij korektorowi | zatwierdź | usuń
SolGeom	**Stanisław Solski** ***Geometra polski*** Jerzy i Mikołaj Schedlowie Kraków 1683	Przepisywany	renata	anna.alfut		0 (0)	0 (0,0str)	nagłówek	odbierz skryptorowi | usuń

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 ☐ żeglarstwo

Poetyka żartu: ☐

Raporty

typ mowy: ---------

rodzaj: literatura faktograficzna

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 ☐ żeglarstwo

poetyka żartu: ☐

region: ---------

50-lecie: 1601-1650

Generuj raport

Znaleziono:

Tekstów: 5

Segmentów: 90891

Wyszukiwarka

Powrót do strony głównej

Wyszukiwarka

Hasło: * korb

Wyszukaj

1 2 3 4

Link
BotŁęczRel IV 84244 korbáczow:
BujnDroga 40801 korbacz/
BystrzInfElem 6926 szkorbutu
BystrzInfRóżn 2263 korbą,
BystrzInfRóżn 2276 korbą:
BystrzInfZup 4499 szkorbutow,
ChmielAteny III 69268 korbé.
ChmielAteny III 80797 korbacz.
CiachPrzyp 7552 skorbućie
DrużZbiór 50587 korbáczem

Reprezentatywność korpusu - możliwości

NKJP

- NKJP ma być reprezentatywny w zakresie recepcji tekstów; w praktyce jest to tylko pewne przybliżenie:
  - arbitralne decyzje dotyczące udziału w korpusie tekstów mówionych i internetowych;
  - wątpliwości dotyczące interpretacji źródeł statystycznych;
  - reprezentatywność tylko pod względem kanału przekazu i stylistyki tekstu.

KORBA

- jeszcze większe trudności w zastosowaniu kryterium reprezentatywności:
  - brak tekstów mówionych;
  - tylko część tekstów barokowych zachowanych do naszych czasów;
  - trudności w pozyskaniu tekstów (rękopisy);
  - mało danych statystycznych o czytelnictwie w epoce;
  - niewielka dostępność tekstów najbardziej popularnych (kalendarze).

Reprezentatywność korpusu - cele

NKJP

- ważnym celem korpusu współczesnego jest odzwierciedlenie świadomości językowej przeciętnego użytkownika danego języka.

KORBA

- w korpusie historycznym ważniejsze jest zgromadzenie tekstów maksymalnie zróżnicowanych pod względem stylistycznym, tematycznym, chronologicznym i regionalnym; dążenie do zachowania reprezentatywności korpusu może utrudniać osiągnięcie tego celu.

Zróżnicowanie tekstów

NKJP

- zapewnienie dużego zróżnicowania stylistyczno-rodzajowego tekstów;
- mniejszą wagę przywiązuje się do zróżnicowania tematycznego, chronologicznego i geograficznego.

KORBA

- dążenie do zapewnienia maksymalnej różnorodności tekstów:
  - stylistyczno-rodzajowa (umożliwia badanie języka poszczególnych gatunków i genrów mowy);
  - tematyczna (umożliwia badanie słownictwa specjalistycznego);
  - chronologiczna (umożliwia badanie zmian językowych);
  - geograficzna (umożliwia badanie języka poszczególnych regionów ówczesnej Polski).

Ortograficzna reprezentacja tekstów

NKJP

- W zasadzie wszystkie teksty zgromadzone w korpusie mają tę samą ortografię (możliwe subtelne różnice, niemające większego wpływu na wyniki wyszukiwania).

KORBA

- Teksty transliterowane zachowują ortografię oryginałów (wysoce niekonsekwentną):
  - konieczność automatycznej transkrypcji i przechowywania w korpusie w dwóch „zrównoleglonych” wersjach.
- Teksty pozyskane z wydań późniejszych mają ortografię XIX- i XX-wieczną:
  - część jest „postarzanych” i będzie miała dwie wersje,
  - część tylko w jednej (współczesnej) ortografii.

Zakres znakowania

Znakowanie w NKJP

- Administracyjne:
  - szczegółowe metadane, ale tylko niewielka ich część jest dostępna dla zewnętrznego użytkownika.
- Strukturalne:
  - ubogie, niedostępne dla użytkownika zewnętrznego;
- Językowe: brak.
- Morfosyntaktyczne: oryginalny, niestandardowy tagset.
- Semantyczne, składniowe, onomastyczne: w podkorpusie 1M.

Znakowanie w KORBie

- Administracyjne:
  - szczegółowe metadane, czyli metryczka tekstu (por. wcześniejszy slajd), wszystkie będą dostepne dla użytkowników.
- Strukturalne: bogate – paginacja, części dzieła, kustosze, metatekst itp.
- Językowe: oznakowanie segmentów z języków innych niż polski.
- Morfosyntaktyczne – problem doboru tzw. tagsetu.
- Semantyczne, składniowe – nieplanowane.
- Onomastyczne: może w podkorpusie 1M.

Znaczniki gramatyczne – porównanie z NKJP

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
miejscownik	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

Znaczniki gramatyczne (cd.) porównanie z NKJP

Osoba: (3 wartości)		
pierwsza	pri	bredzę
druga	sec	bredzisz
trzecia	ter	bredzi
Stopień: (3 wartości)		
równy	pos	cudny
wyższy	com	cudniejszy
najwyższy	sup	najcudniejszy
Aspekt: (2 wartości)		
niedokonany	imperf	iść
dokonany	perf	zajść
niedok./dok	im/perf	aresztować
Zanegowanie: (2 wartości)		
niezanegowana	aff	pisanie, czytanego
zanegowana	neg	niepisanie, nieczytanego

dodane

Znaczniki gramatyczne (cd.) porównanie z NKJP

Akcentowość: (2 wartości)		
akcentowana	akc	jego, niego, tobie
nieakcentowana	nakc	go, -ń, ci
Poprzyimkowość: (2 wartości)		
poprzyimkowa	praep	niego, -ń
niepoprzyimkowa	npraep	jego, go
Akomodacyjność: (2 wartości)		
uzgadniająca	congr	dwaj, pięcioma
rządząca	rec	dwóch, dwu, pięciorgiem
Aglutynacyjność: (2 wartości)		
nieaglutynacyjna	nagl	niósł
aglutynacyjna	agl	niosł-
Wokaliczność: (2 wartości)		
wokaliczna	wok	-em
niewokaliczna	nwok	-m
Kropkowalność: (2 wartości)		
z następującą kropką	pun	tzn.
bez następującej kropki	npun	wg

co z tym?

co z tym?

KORBA a NKJP

- KORBA docelowo ma być podkorpusem NKJP.
- Nie będzie w pełni zgodna metodologicznie.
- Nowy Poliqarp będzie dostosowany do potrzeb KORB-y.

PERSPEKTYWY

- Współpraca z projektami dotyczącymi fleksji historycznej:
  - R. Górski, M. Eder (IJP PAN)
  - M. Woliński et al. (IPI PAN)

Dziękujemy za uwagę!