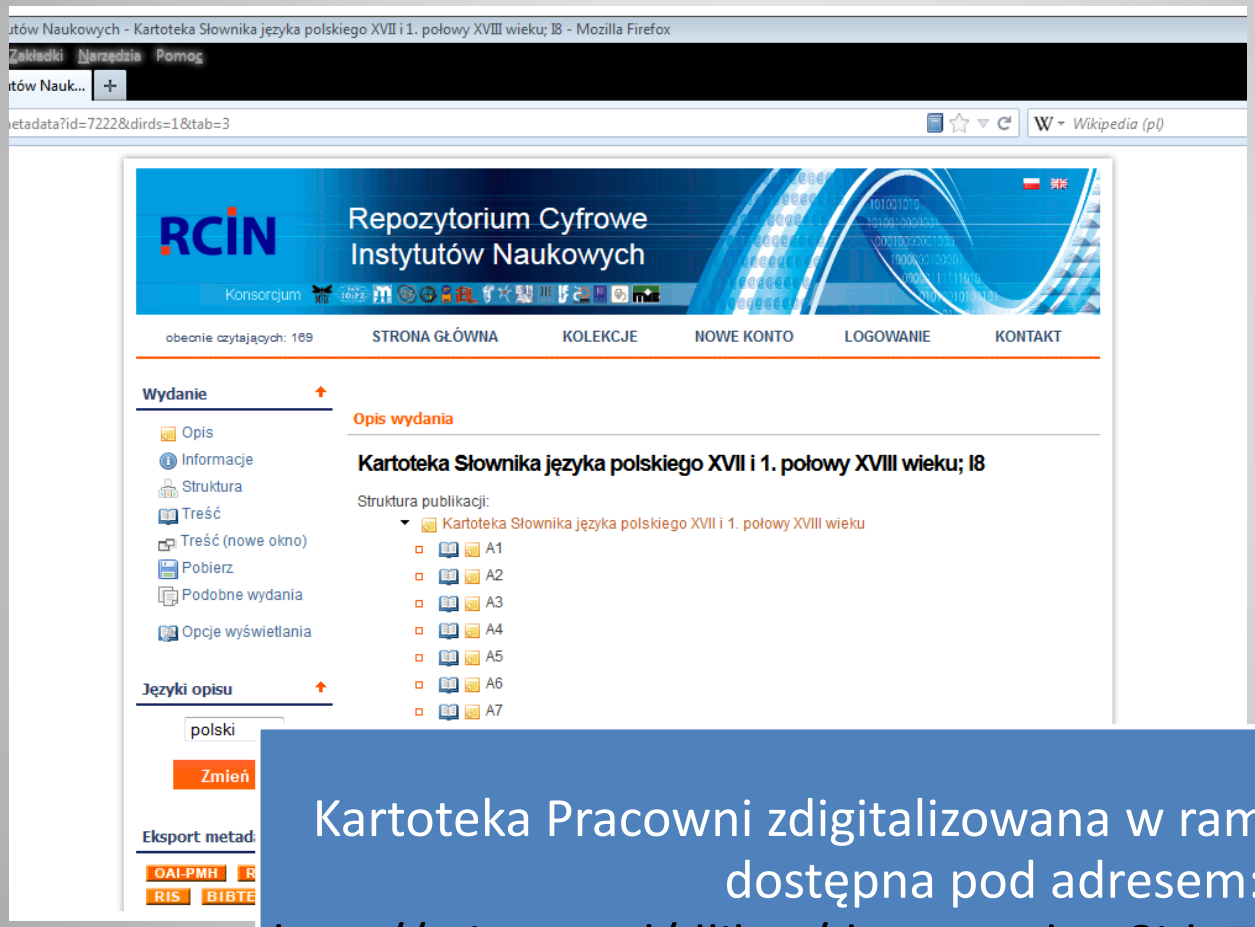
Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.)

Dorota Adamiec – IJP PAN
Włodzimierz Gruszczyński – IJP PAN
Maciej Ogrodniczuk – IPI PAN

Stan przekrojowych badań nad słownictwem polskim XVII-XVIII w.

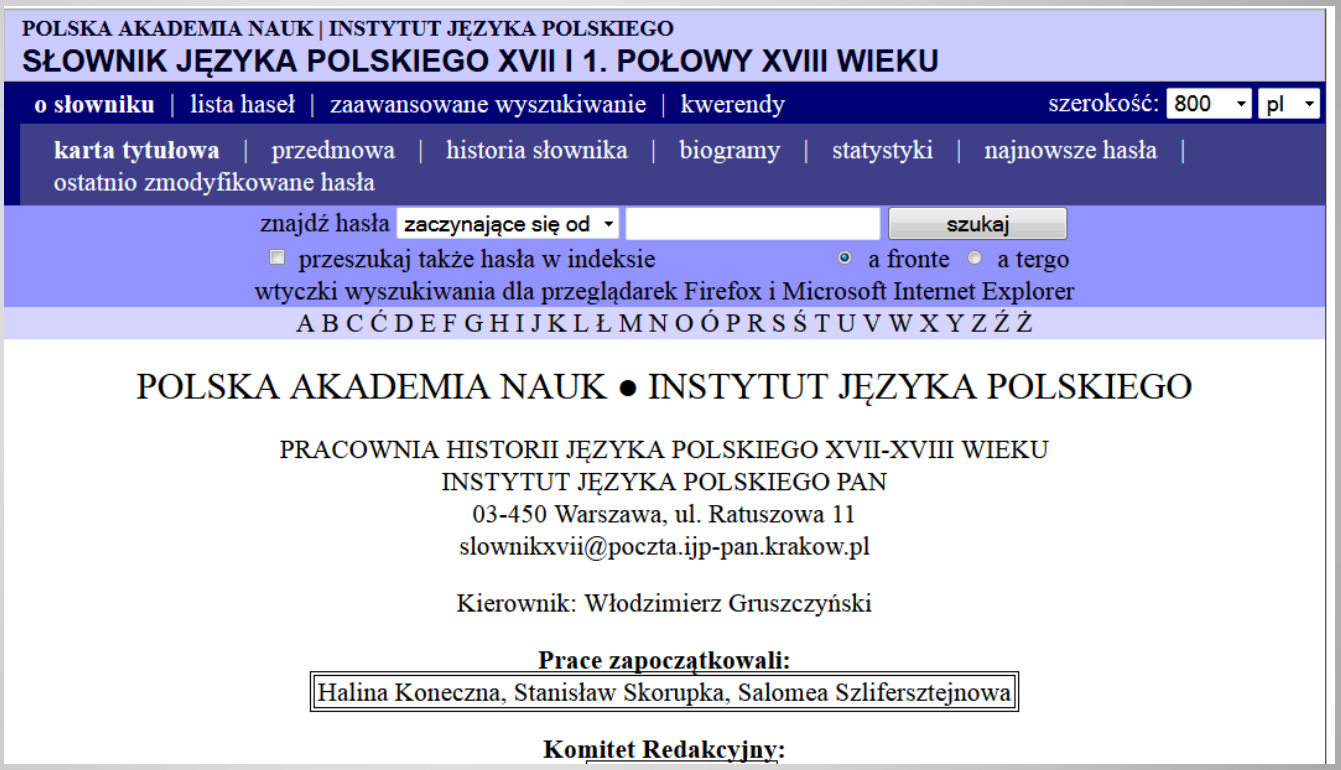
Kartoteka Pracowni Historii Języka Polskiego XVII-XVIII w. (ok. 2,8 mln fiszek)

Stan przekrojowych badań nad słownictwem polskim XVII-XVIII w.



Kartoteka Pracowni zdigitalizowana w ramach projektu RCIN dostępna pod adresem:
http://rcin.org.pl/dlibra/docmetadata?id=7222&dirds=1&tab=3

Stan przekrojowych badań nad słownictwem polskim XVII-XVIII w.



ADRESY:

www.sxvii.pl

xvii-wiek.ijp-pan.krakow.pl/pan_klient/index.php

Konieczne uzupełnienie nowoczesnego warsztatu badawczego – korpus tekstów

- Projekt badawczy „Elektroniczny korpus tekstów polskich XVII i XVIII w. (do 1772 r.)”
- Finansowanie projektu: Narodowy Program Rozwoju Humanistyki na lata: 2013-2017
- Jednostka koordynująca: IJP PAN
- Współpraca: IPI PAN
- Kierownik: Włodzimierz Gruszczyński
- Kryptonim: KORBA (KORpus BArokowy)

Najważniejsze założenia

- Chronologiczne rozszerzenie Narodowego Korpusu Języka Polskiego.
- Planowana objętość: 12 mln segmentów
  - mała w porównaniu z NKJP (300 mln w korpusie zrównoważonym),
  - relatywnie duża jak na korpus historyczny (1 mln w szwedzkich podkorpusach tekstów z okresu Nysvenska, tzn. 1520–1850).

Cele

Dostarczenie nowoczesnego narzędzia historykom języka polskiego:

- wprowadzenie do historii języka metod językoznawstwa korpusowego,
- wsparcie badań historycznojęzykowych (socjolingwistyka, geografia językowa, stylistyka historyczna itp.),
- wsparcie prac nad *Słownikiem języka polskiego XVII i 1 poł. XVIII w.*

Zakres znakowania/tagowania

- znakowanie strukturalne umożliwiające dokładną lokalizację cytatu w tekście:
  - paginacja,
  - rozdziały, części, księgi itp.,
  - notki marginesowe, przypisy itp.,
  - podpisy pod ilustracjami;
- znakowanie socjolingwistyczne (informacje o autorze, wydawcy, tłumaczu);
- znakowanie stylistyczno-genologiczne;
- Znakowanie leksykalne (lematyzacja) – w ograniczonym zakresie;
- Znakowanie morfosyntaktyczne (w podkorpusie 0,5 mln);
- Znakowanie językowe (oznaczenie wszystkich wtrętów obcych, z podziałem na języki);
- Znakowanie typograficzne (ligatury, skróty, być może kroje czcionek itp.).

Zrównoważenie korpusu – problemy

- 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.
- 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 segmentów

  } ok. 1,2 mln segmentów
- Problem dostępności tekstów, możliwości ich pozyskania (np. ograniczone możliwości pozyskiwania rękopisów ze względu na koszto- i czasochłonność).

Gatunki tekstów włączanych do korpusu

- Teksty użytkowe (np.: akta sejmikowe, księgi sądowe, inwentarze, poradniki medyczne i gospodarskie, podręczniki, kalendarze, listy, czasopisma, druki ulotne, teksty specjalistyczne).
- Teksty literackie (np.: diariusze, relacje, historie, herbarze, dyskursy, opowieści, dialogi, dramaty, sielanki, satyry).
- Teksty religijne (np. teksty biblijne, kazania).

Dostępne typy źródeł

- rękopisy,
- starodruki,
- wydania XIX-wieczne i późniejsze,
- wydania współczesne (wydania z nośnika elektronicznego),
- teksty dokładnie transliterowane dostępne w postaci elektronicznej (w szczególności korpus projektu IMPACT – ok. 1,8 mln segmentów)

Rękopisy jako typ źródeł

- Ograniczony zakres możliwości wykorzystania z powodu trudności we wprowadzeniu na nośnik (konieczność czasochłonnego przepisywania przez osobę o wysokich kwalifikacjach);
- Możliwość wykorzystania wydań współczesnych po ich porównaniu z oryginałem;
- Możliwość wykorzystania rękopisów transliterowanych w pracach magisterskich i doktorskich z zakresu edytorstwa lub historii języka polskiego.

Starodruki jako typ źródeł

- Podstawowy typ źródeł w korpusie.
- Większość tekstów będzie przepisywana.
- Część tekstów będzie – być może – skanowana i OCR-owana (za pomocą narzędzia, które powstało w ramach projektu IMPACT).
- Przepisywane będą przede wszystkim teksty udostępnione w bibliotekach cyfrowych.

Wydania XIX-wieczne i późniejsze

- Wydania XIX-wieczne i późniejsze włączymy do korpusu tylko w wyjątkowych wypadkach:
  - jeśli są dostępne w postaci tekstu na nośniku elektronicznym i możliwe jest ich „poprawienie” na podstawie oryginału (rękopisu lub starodruku);
  - jeśli są dostępne w postaci papierowej umożliwiającej wykonanie skanu i OCR, a brak ich podstaw (rękopisy lub starodruki zaginęły).

Wydania współczesne

- Chętnie wykorzystamy wydania współczesne, zwłaszcza jeśli wydawca udostępni tekst na nośniku elektronicznym.
- Konieczność uzyskania praw do wykorzystania (podobnie jak w wypadku pozyskiwania tekstów do NKJP).
- Konieczność porównywania z oryginałami (por. problem z wydaniami IBL).

Znakowanie

- Znakowanie w systemie XML zgodnym ze standardem TEI (w wersji rozszerzonej w trakcie prac nad NKJP)
- Wyzyskany będzie nieco inny podzbiór znaczników (tagset musi zostać dostosowany do swoistych cech tekstów średniopolskich).
- Znakowanie strukturalne, językowe, stylistyczne itp. dokonywane będzie równocześnie z przepisywaniem tekstu dzięki specjalnie zaprojektowanemu formularzowi.

Znakowanie morfosyntaktyczne

- Niewielki podkorpus (ok. 0,5 mln) zostanie oznakowany morfosyntaktycznie.
- Reszta korpusu zostanie oznakowana automatycznie za pomocą tagera stworzonego w ramach projektu.
- Wiarygodność tagowania automatycznego będzie mniejsza niż w wypadku NKJP.

Lematyzacja

- Wszystkie segmenty w korpusie będą rozpoznawane za pomocą lematyzatora, który zostanie dostosowany do prac na tekstach średniopolskich.
- W części korpusu decyzje lematyzatora zostaną skorygowane przez językoznawców.

Problemy (jeszcze) nierozstrzygnięte

- Segmentacja:
  - problem z niekonsekwentną pisownią typu:
    - *Pewnie tak wielkiego stada nieupasiesz albo nadrugie oko musisz olsnąc.*
    - *Mowi Kommendant MSCi Panowie trzeba by nąm tu lęzyka zebys my przecię wiedzieli cosię w Miescie dzieie.*
  - problem ze segmentacją przyjętą w NKJP (odejście od segemtacji ortograficznej).