Dobór tekstów do „Elektronicznego korpusu tekstów polskich z XVII i XVIII w. (do 1772 r.)” – możliwości i ograniczenia budowanego warsztatu badawczego

Dorota Adamiec

Instytut Języka Polskiego PAN

„Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.)” – podstawowe informacje o projekcie

- Projekt badawczy realizowany przez Pracownię Historii Języka Polskiego XVII i XVIII w. Instytutu Języka Polskiego PAN we współpracy z Zespołem Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki PAN.

- Projekt finansowany ze środków Narodowego Programu Rozwoju Humanistyki na lata 2013-2017.

- Kierownik projektu: prof. dr hab. Włodzimierz Gruszczyński.

- Planowana objętość korpusu to 12 mln segmentów.

- Korpus historyczny rozszerzający Narodowy Korpus Języka Polskiego.

Dobór tekstów

- Czynnik decydujący o zawartości korpusu i jego przydatności badawczej.

- Podstawowe założenia autorów korpusu barokowego uwzględniają ogólnie przyjmowane cechy konstytutywne korpusów językowych:
    - reprezentatywność
    - zrównoważenie

Realizacja założeń modelowego korpusu tekstów w korpusie historycznym

- Trudności i ograniczenia:
  - dostępny wyłącznie jeden kanał komunikacyjny – teksty pisane
  - dostępne są tylko teksty, które zachowały się przez 300-400 lat
- Zalety materiału historycznego:
  - zbiór skończony, zamknięty
  - możliwość badawczego oglądu z dużego dystansu czasowego

Dostępne typy źródeł

- rękopisy,
- starodruki (drukowane gotykiem lub antykwą),
- wydania XIX-wieczne i późniejsze tekstów barokowych,
- opracowane elektronicznie wydania współczesne tekstów z XVII i XVIII w.,
- teksty dokładnie transliterowane dostępne w postaci elektronicznej (w szczególności korpus polskiej części międzynarodowego projektu IMPACT – ok. 1,8 mln segmentów).

Płaszczyzny uwzględnione w doborze tekstów do korpusu barokowego

- Zróżnicowanie chronologiczne
- Zróżnicowanie regionalne
- Zróżnicowanie gatunkowe tekstów

Zróżnicowanie chronologiczne

- Zrównoważona ilościowo reprezentacja w korpusie podokresów:
  - 1601-1650
  - 1651-1700
  - 1701-1750
  - 1751-1772
- Cezury czasowe sztuczne – mają charakter wyłącznie porządkujący
- Analizy porównawcze wybranych zjawisk językowych we wskazanych podokresach pozwolą na dostrzeżenie tendencji zmian językowych (rozwojowych i zanikających)

Zróżnicowanie regionalne
(procentowy udział regionów)

	1580-1700 W. Rzepka na podstawie *Bibliografii* Estreichera	1601-0k. 1750 K. Siekierska o źródłach I tomu *Słownika języka polskiego XVII i 1. poł. XVIII w.*
Małopolska	60	24
Wielkopolska	13	13
Mazowsze	8	10
Kresy pn.-wsch.	13	7
Kresy pd.-wsch.	2	16
Pomorze i Prusy	3	5
Śląsk	0,3	2,5
		22% źródeł bez określenia przynależności regionalnej (brak informacji o autorze, teksty anonimowe)

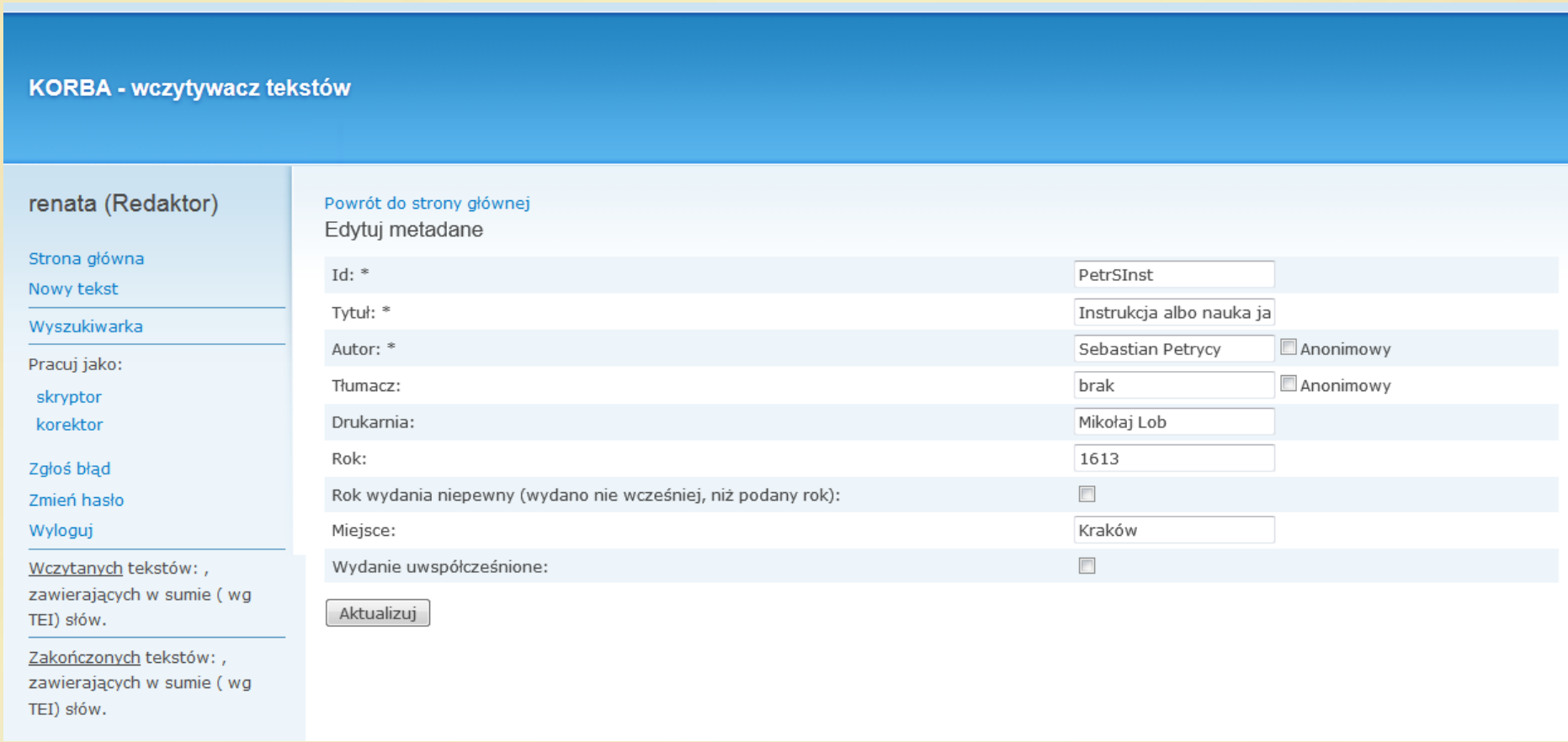
Zróżnicowanie gatunkowe tekstów

- Poezja – znacząca część zachowane go piśmiennictwa z XVII i XVIII wieku – włączana do korpusu w ograniczonym zakresie (archaizmy, neologizmy, indywidualizmy)
- Proza – dążenie do uwzględnienia w korpusie zróżnicowania gatunkowego tekstów reprezentatywnego dla epoki

Zróżnicowanie gatunkowe tekstów prozatorskich

1. Teksty narracyjne: pamiętniki, diariusze, kroniki, historie, relacje podróżnicze, opisy miejsc
2. Proza moralizatorska
3. Proza religijna: teksty biblijne, kazania, postylle, polemiki religijne, żywoty świętych
4. Pisma polityczne
5. Teksty specjalistyczne: traktaty, podręczniki, wzory mów, poradniki, słowniki
6. Teksty urzędowe: akta sejmowe, sejmikowe, sądowe, ordynacje, wilkierze
7. Gazety, druki ulotne
8. Listy

Zestawienie danych o tekście włączanym do korpusu barokowego



Znakowanie

- Korpus budowany z tekstów dobieranych zgodnie z przedstawionymi założeniami jest wielopłaszczyznowo znakowany, aby mógł funkcjonować jako baza danych wyposażona w narzędzia informatyczne.

- Znakowanie tekstów jest wprowadzane w systemie XML zgodnym ze standardem TEI w wersji rozszerzonej na potrzeby NKJP i dostosowanej do cech tekstów średniopolskich.

Znakowanie morfosyntaktyczne

- Niewielki podkorpus (ok. 0,5 mln segmentów) zostanie oznakowany morfosyntaktycznie przez językoznawców.

- Reszta korpusu zostanie oznakowana automatycznie za pomocą stworzonego w ramach projektu tagera (narzędzia informatycznego do automatycznej anotacji morfoskładniowej).

- Wiarygodność tagowania automatycznego będzie mniejsza niż w wypadku NKJP (ze względu na znaczną wariantywność i brak stabilizacji gramatycznej w języku średniopolskim).

Oczekiwane korzyści badawcze w zakresie analizy gramatycznej polszczyzny barokowej

- Opracowanie całościowego opisu fleksyjnego tego okresu rozwojowego polszczyzny.
- Wskazanie tendencji rozwojowych na przestrzeni prawie 200 lat.
- Możliwość weryfikacji dotychczasowych ustaleń historyków języka baroku opartych na mniejszych próbach materiałowych.