

Projekt zintegrowanego systemu informatycznego do studiowania polszczyzny XVII i XVIII wieku



Włodzimierz Gruszczyński
Instytut Języka Polskiego PAN

wlodzimierz.gruszczyński@ijp.pan.pl

Maciej Ogrodniczuk
Instytut Podstaw Informatyki PAN

maciej.ogrodniczuk@ipipan.waw.pl

Finansowanie projektu



NARODOWY PROGRAM
ROZWOJU HUMANISTYKI

Prezentacja powstała w ramach projektu pt.

Rozbudowa Elektronicznego Korpusu Tekstów Polskich XVII i XVIII w. i jego integracja z *Elektronicznym słownikiem języka polskiego XVII i XVIII w.*

finansowanego z funduszy Narodowego Programu Rozwoju Humanistyki MNiSW i realizowanego w latach 2019–2023 pod kierownictwem W. Gruszczyńskiego w IJP PAN we współpracy z Zespołem Inżynierii Lingwistycznej IPI PAN.

Założenia projektowanego systemu

1. Cel: stworzenie platformy informatycznej ułatwiającej studiowanie polszczyzny XVII i XVIII wieku.
2. Zasoby udostępniane obecnie odrębnie:
 - a. *Elektroniczny słownik języka polskiego XVII i XVIII wieku (e-SXVII)*,
 - b. Elektroniczny Korpus Tekstów Polskich z XVII i XVIII Wieku (KorBa),
 - c. Cyfrowa Biblioteka Druków Polskich i Polski Dotyczących z XVI, XVII i XVIII Wieku (CBDU),
 - d. *Kartoteka Słownika języka polskiego XVII i 1. poł. XVIII w.*
3. Narzędzia:
 - a. wyszukiwarki korpusowe,
 - b. wyszukiwarki tekstowe,
 - c. Chronofleks.

Dwa kierunki integracji zasobów historycznojęzykowych

- zasoby tego samego typu łączone na osi czasu (np. korpus diachroniczny),
- zasoby różnego typu łączone w obrębie tego samego okresu.

	przed XVI w.	XVI w.	XVII-XVIII w.	XIX w.	XX-XXI w.
słowniki			e-SXVII		
korpusy			KorBa		
biblioteki cyfrowe			CBDU		
kartoteki i in. zasoby			Kartoteka SXVII		

Zasoby podlegające integracji (1)



Elektroniczny słownik języka polskiego XVII i XVIII wieku

- dostęp: <https://sxvii.pl/> lub https://xvii-wiek.ijp.pan.pl/pan_klient/;
- zawartość: ok. 39 tys. artykułów hasłowych na różnym stopniu opracowania (w tym ok. 25 tys. załączków artykułów);
- formy hasłowe i in. formy fleksyjne (jedynie poświadczone) podawane w transkrypcji;
- cytaty dokumentacyjne podawane w kolejności chronologicznej w transliteracji;
- brak informacji statystycznej (ze względu na zmieniającą się podstawę materiałową).

Zasoby podlegające integracji (2)

Elektroniczny Korpus Tekstów Polskich z XVII i XVIII Wieku

- dostęp: <http://korba.edu.pl/>;
- zawartość:
 - obecnie 13,5 mln segmentów (ok. 700 tekstów z lat 1601–1772),
 - docelowo 25 mln segmentów (w tym ok. 4 mln z lat 1773–1800);
- teksty dostępne zarówno w transkrypcji, jak i transliteracji;
- teksty oznakowane morfosyntaktycznie przez tagery Concraft i Toygger;
- podkorpus o wielkości 0,5 mln segmentów oznakowany ręcznie;
- bardzo dokładne metadane źródeł;
- możliwości wyszukiwania wg różnych kryteriów.

Zasoby podlegające integracji (3)

Cyfrowa Biblioteka Druków Polskich i Polski Dotyczących z XVI, XVII i XVIII Wieku

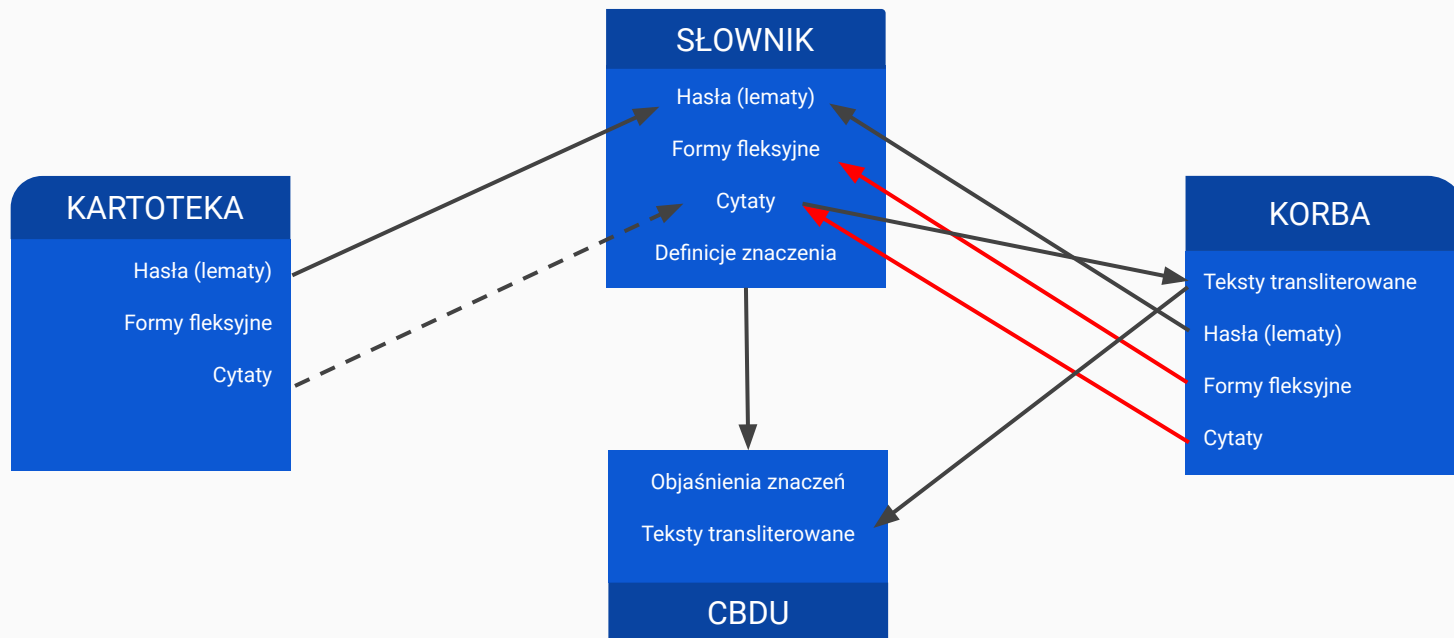
- dostęp: <https://cbdu.ijp.pan.pl/>;
- zawartość:
 - ok. 1500 skanów starodruków w 11 językach, w tym 326 w języku polskim,
 - bardzo bogate metadane, w tym informacje o relacjach łączących dany tekst z innymi (przeróbki, tłumaczenia);
 - słowniczki trudniejszych słów i zwrotów (przy tekstach polskich),
 - objaśnienia niektórych słów i zwrotów łacińskich;
 - komentarze historyczne dotyczące faktów opisywanych w tekstach.

Zasoby podlegające integracji (4)

Kartoteka *Słownika języka polskiego XVII i 1. poł. XVIII w.*

- dostęp: <http://rcin.org.pl/dlibra/publication?id=20029&tab=3>;
- zawartość: skany ok. 1,8 mln rękopiśmiennych kart cytatowych powstałych w latach 1955–2011 w wyniku ręcznej ekscerpcji źródeł z XVII i połowy XVIII w.;
- karty ułożone alfabetycznie wg form hasłowych wyróżnionych w cytacie wyrazów;
- część cytatów jest niekompletna;
- niektóre lematyzacje dyskusyjne.

Planowane i **wprowadzone** powiązania między zasobami



Wdrożone powiązania e-SXVII ← KorBa

1. Automatyczne wyszukiwanie wystąpień form wyrazu hasłowego w korpusie

The screenshot shows the e-SXVII online dictionary interface. At the top left is the logo 'eSXVII'. To the right are links for 'o słowniku' and 'kwerendy', and flags for Polish and English. The search term '*DUBITOWAĆ' is entered, with a 'czas. ndk' tag. A status bar indicates 'W TRAKCIE OPRACOWANIA' with a printer icon. A list of categories is shown with expandable sections: 'Notowanie w słownikach', 'Formy gramatyczne', 'Etymologia', 'Znaczenia', and '»mieć wątpliwości, powątpiewać«'. The 'Znaczenia' section is expanded, showing a definition: '– Iak tedy Obaczyli ze Ięm Smaczno y pię y rzeswość dopiero uwierzyli zem sanus mente et Corpore [zdrów na umyśle i ciele] oczym Cyrulik osobiwie bardzo dubitował.' followed by a citation 'PasPam 280-280v.'. A red speech bubble with the text 'KLIK!' points to the 'Więcej cytatów w Korpusie Barokowym' link at the bottom of the expanded section.

eSXVII

o słowniku kwerendy

*DUBITOWAĆ czas. ndk

W TRAKCIE OPRACOWANIA

▼ Notowanie w słownikach

▼ Formy gramatyczne

▼ Etymologia

▼ Znaczenia

▼ »mieć wątpliwości, powątpiewać«

- Iak tedy Obaczyli ze Ięm Smaczno y pię y rzeswość dopiero uwierzyli zem sanus mente et Corpore [zdrów na umyśle i ciele] oczym Cyrulik osobiwie bardzo dubitował. [PasPam 280-280v.](#)

▼ Więcej cytatów w Korpusie Barokowym

KLIK!

Wdrożone powiązania e-SXVII ← KorBa

1. Automatyczne wyszukiwanie wystąpień form wyrazu hasłowego w korpusie

KORBA INFORMACJE ▾ INSTRUKCJA LOGOWANIE

ELEKTRONICZNY KORPUS TEKSTÓW POLSKICH Z XVII I XVIII W. (DO 1772 R.)

Korpus
Korba automatyczna ▾

Zapytanie
((base="dubitować"&!pos="ger|ppas|pact|ppraet"))

á é Á Ę

KONSTRUKTOR ZAPYTAŃ ODRZUĆ OBCE SEGMENTY METADANE ▾

Warstwa wyświetlania
uwspółcześniona ▾ Liczba wyników na stronie
10

Wyszukaj

Znaleziono 20 wyników.

20 dodatkowych cytatów z różnych źródeł.

Lp	Lewy kontekst	Rezultat	Prawy kontekst	Skrót	Data
1	tej sentencji pozorniejszej o Relikwiach, wyliczonych Świętych mozem nie	dubitować [dubitować:inf:imperf]	, ale ich wenerować bez szkrupulu. W których miejscach	ChmielAteny_III	1754
2	Głowę Z. był spektatorem cudu, bo Malchus Zakonnik	dubitujący [dubitować:pact:sg:nom:m:imperf:aff:pos]	aby Głowa Z. Jana mogła się przez 760.	ChmielAteny_III	1754

Wdrożone powiązania e-SXVII ← KorBa

2. Automatyczne wyszukiwanie dodatkowych poświadczeń form fleksyjnych w korpusie

The screenshot shows the e-SXVII online dictionary interface. At the top left is the logo 'eSXVII'. To the right are links 'o słowniku' and 'kwerendy', and flags for Polish and English. Below the header is a navigation bar with letters A-Z and Polish characters. The main content area displays the word 'BACHMAT' with grammatical tags 'rzecz.' and 'm'. To the right, it says 'W TRAKCIE OPRACOWANIA' with a printer icon. There are two expandable sections: 'Notowanie w słownikach' and 'Formy gramatyczne'. The 'Formy gramatyczne' section is expanded, showing a list of grammatical forms for 'bachmat' and 'bachmaty' in various cases and numbers, each with a small icon.

eSXVII

o słowniku kwerendy

ABCĆDEFGHIJKLMNOPRÓQSSTUVWXYZŹŻ

BACHMAT rzecz. m

W TRAKCIE OPRACOWANIA

Notowanie w słownikach

Formy gramatyczne

lp M. bachmat
D. bachmata
B. bachmata
N. bachmatem
Ms. bachmacie
W. bachmacie

lm M. bachmaty
D. bachmatów
B. bachmaty
N. bachmaty
Ms. bachmatach

Braki we wdrożonych powiązaniach

- 1) Brak możliwości wyszukiwania form fleksyjnych niewykazanych w artykule hasłowym e-SXVII.
- 2) Problemy z wyszukiwaniem form fleksyjnych wariantów fonetycznych hasła.
- 3) Brak interfejsu dla autorów artykułów hasłowych w e-SXVII.

Eksperymentalne powiązania e-SXVII → CBDU

Cel:

wykorzystanie definicji znaczeń zawartych w słowniku historycznym do objaśniania znaczeń wyrazów (archaizmów) wskazanych przez czytelnika w tekstach udostępnianych w bibliotece cyfrowej.

Eksperymentalne powiązania e-SXVII → CBDU

Problemy:

1. Konieczność utworzenia tekstowej wersji dokumentu dostępnego w postaci graficznej (pdf) i “podłożenia” jej pod wersję graficzną (skanowanie? przepisywanie?).
2. Konieczność lematyzacji każdej z form tekstowych, aby możliwe było odnalezienie odpowiedniego artykułu hasłowego w e-SXVII.
3. Wybór haseł (lematów), które są potencjalnie trudne dla czytelników (nieuwzględnianie haseł występujących w słowniku współczesnego języka polskiego? zmiana znaczenia hasła?)
4. Wybór definicji w wypadku haseł wieloznacznych (czy to w ogóle możliwe?)

Eksperymentalne powiązania e-SXVII → CBDU

się Wojsko, Kawaleryą y Infanteryą która ielzcze po wieczce przeszley nie weszła była, wprowadził do Miastá; potym rzęsiłto z Dział strzeląc zaczął, tak iż kilka koni pod temi zabito, którzy byli przy Krolu I. Mći.

Die 6. ejusdem. Ruszył Krol I. Mć z całym Wojskiem, y zbliżył się ku Miastu, tak blisko, iż Kwartyerá Iego Krolewskiej Mći tylko iest o ćwierć mile od Miastá. Iuz tedy Wojsko Iego Krolewskiej Mći opasało Rygę, że nikt ná ląd wyniść nie może. Wzięto przy tym Insulę Lutzáverholm, z ktorey dobrze do Miastá szturmować może.

Die 7. Káprowie podsunęli się byli pod Kwartyerę Iego Krolewskiej Mći, y poczęli ognia dawać, lecz od Armát Iego Krolewskiej Mći są spędzeni, *nec amplius comparent.*

(5)

Die

eSXVII

INSUŁA (rzech. ż)
»wyspa«

INSUŁA (rzech. ż)
»wyspa«

Próby powiązania kartoteki SXVII z e-SXVII

Cel:

Stworzenie narzędzia ułatwiającego autorom artykułów hasłowych w e-SXVII odnajdowania fiszek odnoszących się do opracowywanego artykułu hasłowego.

Problem:

Niemożliwe jest zastosowanie programów do OCR, ponieważ fiszki są w większości rękopiśmienne, robione przez różne osoby.

BUDYNEK

ertowie z rozumem, u budynkach
fis kocha, budowaniem fis zabawia

MTodrz Karz Hom 1681, 202/111

Próby powiązania kartoteki SXVII z e-SXVII

Eksperyment:

Podjęto próbę mającą na celu zindeksowanie (fragmentu) kartoteki, tzn. ustalenia dla każdego reprezentowanego w niej hasła adresu pierwszej fiszki odnoszącej się do danego hasła. Jeśli eksperyment się powiedzie, zindeksowana zostanie cała kartoteka i dzięki temu redaktor słownika (być może także jego czytelnik) będzie mógł uzyskać dostęp do odpowiedniego fragmentu kartoteki za pomocą jednego kliknięcia.

Dziękujemy za uwagę.