KorBa

to iest

Elektroniczny Korpus Textow Polskich z XVII i XVIII w. (do 1772 r.)

W Wárßáwie, dniá 18. Iunii A.D. MMXVIII

Podstawowe dane projektu

- Tytuł: ***Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.)***
- Kryptonim: *KorBa* (**Kor**pus **Ba**rokowy).
- Finansowanie: Narodowy Program Rozwoju Humanistyki MNiSW.
- Okres finansowania: 26.03.2013-25.03.2018.
- Budżet: 900 000 PLN.
- Jednostka koordynująca: Instytut Języka Polskiego PAN.
- Współpraca: Zespół Inżynierii Lingwistycznej IPI PAN.
- Planowana objętość: 12 mln segmentów.
- Planowana objętość podkorpusu znakowanego ręcznie: 0,5 mln segmentów.
- Planowana forma udostępniania: on-line
  - dla użytkowników zalogowanych – bez ograniczeń,
  - dla użytkowników niezalogowanych – niewielkie ograniczenia ilościowe (albo przeszukiwanych źródeł, albo udostępnianych wyników).

Najważniejsze założenia projektu

- Różnorodność tekstów zapewniająca jak największą reprezentatywność chronologiczną, geograficzną, genologiczną, tematyczną itp.
- Bogate metadane powiązane z tekstami:
  - dane bibliograficzne,
  - charakterystyka stylistyczno-genologiczna,
  - charakterystyka tematyczna,
  - charakterystyka socjolingwistyczna i geolingwistyczna.
- Dostęp do transliteracji i transkrypcji.
- Szczegółowe znakowanie struktury tekstu.
- Znakowanie wszystkich wtrętów obcych, z podziałem na języki.
- Oznakowanie morfosyntaktyczne (w części ręczne, w części – automatyczne).

Charakter projektu

- Projekt miał charakter heterogeniczny.
- Zadania o charakterze filologicznym:
  - wybór reprezentatywnych tekstów z epoki,
  - opracowanie zasad transliteracji i transkrypcji,
  - przeniesienie tekstów na nośnik elektroniczny,
  - opracowanie systemu znaczników różnego rodzaju,
  - opatrzenie tekstów szczegółowymi metadanymi itp.
- Zadania o charakterze informatycznym:
  - stworzenie narzędzi informatycznych służących do gromadzenia, przetwarzania, przeszukiwania i prezentowania tekstów zawartych w korpusie,
  - modyfikacja narzędzi już istniejących, stworzonych na potrzeby korpusów tekstów współczesnych.

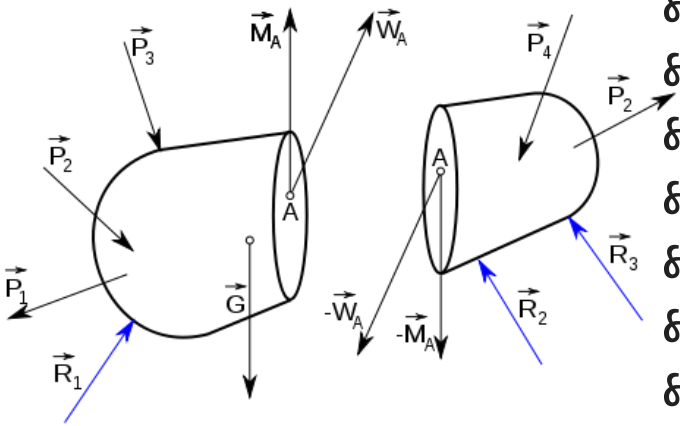
Zespół – 3 x K

- **K**ierownik projektu:
  - Włodzimierz Gruszczyński
- **K**oordynatorka:
  - Renata Bronikowska
- Prace **k**oncepcyjne:
  - Dorota Adamiec
  - Anna Andrzejczuk
  - Renata Bronikowska
  - Włodzimierz Gruszczyński
  - Witold Kieraś
  - Małgorzata B. Majewska
  - Maciej Ogrodniczuk
  - Adam Przepiórkowski
  - Marcin Woliński

Zespół – skryptorzy

- Anna Alfut
- Agnieszka Adamiec
- Dorota Adamiec
- Anna Arendt
- Alicja Bielak
- Laura Bielak
- Renata Bronikowska
- Katarzyna Bury
- Maciej Bury
- Jolanta Czajkowska
- Julia Domitrak
- Mirella Gliwińska
- Zuzanna Głuszczak
- Jolanta Gomółka
- Izabela Jagielska
- Klaudia Jovanovska
- Ewa Karasińska-Gajo
- Magdalena Kołodziejczyk
- Wojciech Kordyzon
- Anna Krasowska
- Katarzyna Kryńska
- Agnieszka Łodzińska
- Małgorzata Maciejewska
- Magdalena Majdak
- Olga Makarova
- Ewelina Mantorska
- Emanuel Modrzejewski
- Wiesław Morawski
- Małgorzata Pachulska
- Aldona Przyborska-Szulc
- Paweł Siemieniak
- Dawid Siwicki
- Paulina Wdowska
- Emilia Zdankiewicz
- Anna Żółtak

Zespół – anotatorzy



- Dorota Adamiec
- Renata Bronikowska
- Włodzimierz Gruszczyński
- Piotr Janas
- Matylda Kozłowska
- Dawid Lipiński
- Magdalena Majdak
- Emanuel Modrzejewski
- Wiesław Morawski
- Izabela Pawlak
- Ewelina Pędzich
- Marcin Podlaski
- Aldona Przyborska-Szulc
- Ewa Rodek
- Paulina Rosalska
- Aleksandra Wieczorek
- Sebastian Żurowski

Zespół – inne prace filologiczne

- Dorota Adamiec
- Renata Bronikowska
- Włodzimierz Gruszczyński
- Witold Kieraś
- Monika Kresa
- Paweł Kupiszewski
- Małgorzata B. Majewska
- Wiesław Morawski
- Aldona Przyborska-Szulc

Prace pomocnicze – indeks SXVII

- Mateusz Adamczyk
- Ewa Balicka
- Dagmara Banasiak
- Agata Hącia
- Aleksandra Wieczorek

Zespół – informatycy, programiści

- Bartłomiej Borek
- Zbigniew Gawłowicz
- Piotr Janas
- Witold Kieraś
- Łukasz Kobyliński
- Jakub Kostrzewa
- Dorota Komosińska
- Katarzyna Krasnowska-Kieraś
- Michał Lenart
- Maciej Ogrodniczuk
- Katarzyna Streich
- Jan Szejko
- Michał Wasiluk
- Aleksander Zabłocki
- Bartosz Zaborowski
- Mateusz Żółtak

Instytucje i osoby, które nas wspierały

- Biblioteka Wolne Lektury
- Cyfrowa Biblioteka Narodowa Polona
- Zespół Bibliotek Cyfrowych Poznańskiego Centrum Superkomputerowo-Sieciowego
- prof. dr hab. Janusz S. Bień
- Iwona Maciejewska
- prof. dr hab. Wacław Walecki i wydawnictwo Collegium Columbinum
- Katarzyna Zawilska
- Wojciech Żółtak

Dziękujemy!

Obecny stan korpusu

- 718 tekstów o łącznej długości 10,97 mln słów, co daje prawie 13,5 mln segmentów w rozumieniu NKJP.
- Wszystkie teksty mają dokładne metryczki.
- Wszystkie teksty mają oznakowaną strukturę i oznaczone słowa w językach obcych.
- Niektóre teksty wymagają ciągle korekty ☹.
- Działają następujące narzędzia:
  - Wczytywacz,
  - Transkryber,
  - Anotatornia 2,
  - Korbeusz, czyli postarzony Morfeusz,
  - Wyszukiwarka MTAS (zmodyfikowana).

System znaczników morfosyntaktycznych, to jest TAGSET

Klasy fleksyjne – różnice w stosunku do NKJP

	NKJP			KorBa		
Leksem	**Fleksem**	**Ozn.**	**Przykład**	**Fleksem**	**Ozn.**	**Przykład**
Rzeczownik	● rzeczownik ● forma depr.	subst depr	*królowie* *króle*	Rzeczownik	subst	*królowie, wilcy, anioły, anieli*
Liczebnik przymiotnikowy	● –			liczebnik przymiotnikowy	adjnum	*drugi, dwukrotny, dwojaki, samowtóry*
Liczebnik przysłówkowy	● –			liczebnik przysłówkowy	advnum	*dwukrotnie, dwojako, dwakroć, samowtór*
Przymiotnik	● przymiotnik ● przymiotnik poprzyimkowy ● przymiotnik przyprzymiot. ● przymiotnik predykatywny	adj adjp adja adjc	*polski* *polsku* *polsko* *wesół*	● przymiotnik ● przymiotnik odm. niezłożona ● przymiotnik przyprzymiot.	adj adjb adja	*dobry, zdrowego; zdrów, gwałtownę, polsku;* *angielsko, ziemno*
Czasownik	● – ● – ● – ● – ● –			● forma przeszła BYĆ ● forma BYĆ jako składnik cz. przyszłego ● aglutynant aoryst. ● imiesłów przym. czyn. odmiana niezłożona ● imiesłów bierny odmiana niezłoż.	plusq fut agltaor pactb ppasb	*był* *będzie* *-(e)ch, -(e)chmy* *będęcy, jadęcy* *umęczon, ukrzyżowan*

Znaczniki gramatyczne – różnice w stosunku do NKJP

- Wprowadzenie liczby podwójnej (`du`).
- Zmiany w nazwach i kryteriach ustalania podrodzajów męskich (`manim1` i `manim2`).
- Wprowadzenie rodzajów przymnogich (`p1`, `p2`)
- Wprowadzenie trzeciej wartości kategorii aspektu (`biasp`).
- Rezygnacja z kategorii akomodacyjności form liczebników.

Kategorie gramatyczne: liczba

Liczba: (3 wartości)		
pojedyncza	sg	niewiasta
podwójna	du	M. B. W. (dwa) szczyta, męża; (dwie) robocie, ręce, świecy; (dwie) ście, plecy D. Msc. (dwu) panu, królu; (dwu) kopu, niedzielu; (dwu) latu, pokoleniu C.N. (dwiema) mężoma, zakonoma; (dwiema) niewiastama, rzeczoma; (dwiema) latoma, plecoma
mnoga	pl	niewiasty

- Wszystkie formy dawnej liczby podwójnej znakujemy jako du. Nawet jeśli forma rzeczownika o końcówce typowej dla liczby podwójnej łączy się z przymiotnikiem w liczbie mnogiej, to i tak uznajemy ją za formę *dualis*, np. we frazie *(robił to) swoimi starymi rękoma* formę *rękoma* znakujemy jako du:inst, chociaż przymiotnik *starymi* znakujemy jako pl:inst:f:pos.

Kategorie gramatyczne: rodzaj

- W ustalaniu rodzaju rzeczowników w języku średniopolskim nie mogą pomóc konteksty diagnostyczne, które stosowane były w czasie anotacji NKJP, ponieważ anotatorzy nie mają kompetencji językowej w zakresie języka średniopolskiego.
- Przypisujemy rodzaj **poszczególnym formom**, nie zwracając uwagi na to, że w konsekwencji różne formy potencjalnie tego samego leksemu mogą mieć przypisany różny rodzaj.
- Jeśli byłyby to formy rzeczywiście tego samego leksemu, to wartości kategorii rodzaju, które im zostały przypisane, muszą być niesprzeczne (uzgadnialne), np.: `m` i `mnanim` albo `f.n.p1.p2` i `n`.

Kategorie gramatyczne: rodzaj

Rodzaj przypisujemy poszczególnym formom, nie zawsze znając formę podstawową. Tryb przypisywania rodzaju powinien był być następujący:

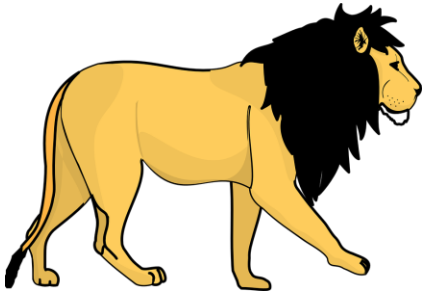
- formom leksemów znanych anotatorowi i zgodnym ze współczesną normą – przypisujemy rodzaj wg intuicji;
- formom, do których są różne podpowiedzi, przypisujemy wartość wspólną, np. formie *alarmami*, na podstawie podpowiedzi z lematami ALARM, ALARMA, ALARMO przypisujemy wartość rodzaju `0`, a lemat `alarm*: subst:pl:inst:0`;
- formom nieznanych leksemów, do których jest jedna podpowiedź, przypisujemy interpretację z podpowiedzi.

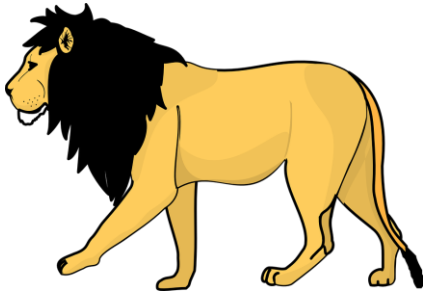
Rodzaj

Rodzaj		
nieznany/ wspólny	m.f.n.p1.p2	(tymi) narańczagami (lemat NARAŃCZAG*) (tymi) rożynkami (lemat ROŻYN*) (o tych) franbugach (lemat FRANBUG*)
niemęski	f.n.p1.p2	(tych) suden (lemat SUDN*)
nieżeński	m.n	(na tym) darniu (lemat DAR*)
nienijaki	m.f	(to jest) adwena (lemat ADWENA)
	m.n.p2	emolumenta (lemat EMOLUMENT*)
męski uogólniony	m	(to jest) detryment (lemat DETRYMENT) (temu) forytarzowi (lemat FORYTARZ) indycentowi (lemat INDYCENT)
męski nieżywotny	mnanim	stół, dom, cebrzyk (nie widział) purgansu (lemat PURGANS) (na) purgans (lemat PURGANS)
	manim.p2	(widzę) temperamenta (lemat TEMPERAMENT*)
męski żywotny	manim	baranek, babsztyl, (to był) assawoła (lemat ASSAWOŁA)
męski żywotny jak osob.	manim1	(srodzy) tygrysowie (lemat TYGRYS) (groźni) narodowie, inszy narodowie planetowie (lemat PLANET*) (wszytkie zwierzęta lwi i niedźwiedzie, pardzi, psi, smocy (lematy: LEW, NIEDŹWIEDŹ, PARD*, PIES, SMOK) (widzi) subalternów (lemat SUBALTEREN*) (prowadzono) słoniów
męski żywotny jak nieosob.	manim2.p2	(zwojował) Węgry i Pomorczyki , (korzenić) schysmatyki i heretyki,
żeński	f	stuła, kobiety, (przeciw) anginie (lemat ANGINA) na pułnocney eluardzie (lemat BELUARDA)
nijaki	n	dziecko, okno, co, (uczynili) bando (lemat BANDO)
przymnogi osob.	p1	(jadą owi) Państwo, królestwo (jedli)
przymnogi	p2	(grać w) arcaby (lemat ARCABY)

Kategorie gramatyczne: rodzaj męski żywotny – założenia opisu

- Wiele rzeczowników rodzaju manim może mieć dwie konkurencyjne formy `pl:nom`, np.:
  - *(te) lwy, ptaki, anioły, pany, chłopy,*
  - *(ci)* ***lwowie****, ptacy, anieli||**aniołowie**, panowie, chłopi.*
- Formom tym przypisujemy odpowiednio rodzaj `m` i `manim1`.
- Jeżeli w `pl:nom` następuje neutralizacja tego typu form, np. (*ci* i *te*) *gospodarze*, formę tekstową znakujemy jako `pl:nom:m`, chyba że uzgadnia się z formą w rodzaju `manim1`, np. z formami typu *ci*, *wielcy*, *byli*.





Kategorie gramatyczne: rodzaj męski żywotny – praktyka anotacyjna

- Forma *wilcy* w podkorpusie ręcznie anotowanym wystąpiła 7 razy.
  - Rodzaj `manim1` został jej przypisany 1 raz;
  - Rodzaj `m` został jej przypisany 6 razy.
- Forma *wilcy* w podkorpusie anotowanym automatycznie wystąpiła 92 razy.
  - Rodzaju `manim1` tager nie przypisał ani razu.
  - We frazie ***wilcy drapieżni*** forma przymiotnika jest systematycznie znakowana jako `manim1`. ☹

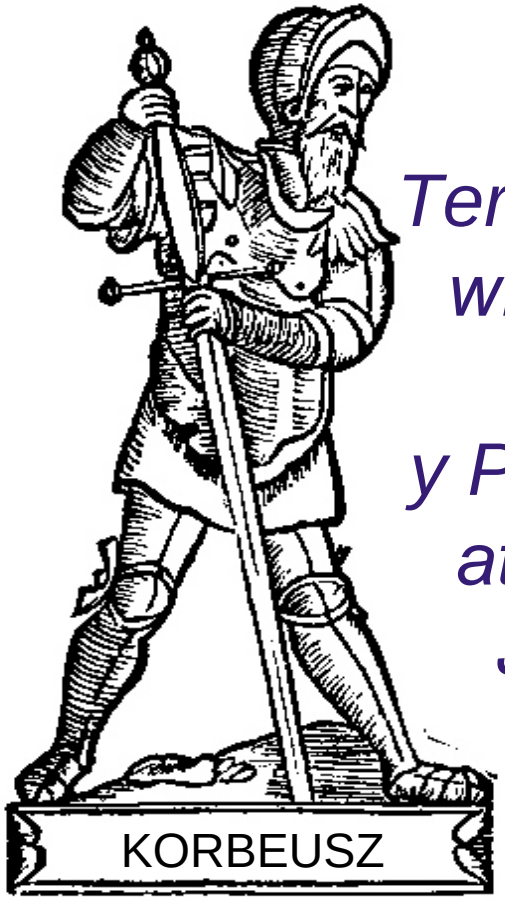
KorBa2

Elektroniczny Korpus Tekstów Polskich XVII i XVIII w. – perspektywy

- W marcu 2018 r. złożony został powtórnie wniosek o sfinansowanie kontynuacji projektu przez NPRH.
- Tytuł projektu: Rozbudowa Elektronicznego Korpusu Tekstów Polskich XVII i XVIII w. i jego integracja z *Elektronicznym słownikiem języka polskiego XVII i XVIII w.*
- Okres finansowania: 2019-2024.
- Jednostka koordynująca: IJP PAN.
- Kryptonim: KorBa2.
- Współpraca: Zespół Inżynierii Lingwistycznej IPI PAN.
- Zespół w znacznym stopniu ten sam, co w I etapie.

Cele nowego projektu:

- Rozbudowa Elektronicznego Korpusu Tekstów Polskich XVII i XVIII w. (do 1772 r.) utworzonego w I etapie projektu i przekształcenie go w Elektroniczny Korpus Tekstów Polskich XVII i XVIII w.:
  - uzupełnienie korpusu o teksty od 1772 r. do końca XVIII w.,
  - dodanie nowych tekstów do korpusu już istniejącego:
  - ważne teksty, które zostały włączone we fragmentach,
  - zwiększenie liczby tekstów z początków XVIII w.,
  - większe zrównoważenie stylistyczne.
- Łączne powiększenie objętości korpusu o 12 mln segmentów.
- Udoskonalenie narzędzi informatycznych obsługujących korpus:
  - transkrybera,
  - analizatora morfologicznego (dostosowanie tagsetu do tekstów z końca XVIII w.),
  - tagera.
- Integracja korpusu z *Elektronicznym słownikiem języka polskiego XVII i XVIII w.* – stworzenie narzędzi do automatycznej ekstrakcji danych i wspomagających wybór najlepszych przykładów.
- Częściowa integracja z NKJP – stworzenie możliwości przeszukiwania materiału współczesnego i z innych epok.
- Przeprowadzenie analiz lingwistycznych materiału zawartego w korpusie: syntaktycznych, fleksyjnych i semantyczno-leksykalnych.



Teraz przy zákończeńiu tey przemowy wßyscy authorowie, vniżony pokłon oddáiąc, J.W.W.Mćiom Pańiom y Panom za ich wysłuchánie z gorącą attencyą pokorńie dźiękuią y życzą J.W.W.Mćiom zupełney obfitośći wßelákiego ßczęśćia.

W Wárßáwie, dies 18 Iunii A.D. CIↃCIↃXVIII