



Korpus historyczny, czyli problemata transliteracji, transkrypcji i annotacji

na przykładzie

Elektronicznego Korpusu Textów Polskich
z XVII i XVIII w. (do 1772 r.)

metodą wokalną wyłożone i rycinami zexemplifikowane przez:

Włodzimierza Gruźczyńskiego,

Dorotę Adamiec,

Renatę Bronikowską

Instytut Języka Polskiego PAN

Witolda Kierafia,

Dorotę Komosińską

✉ Marcina Wolińskiego

Instytut Podstaw Informatyki PAN



W Warszawie, dnia 23 Aprila A.D. MMXVIII

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



- ❧ Finansowanie projektu: Narodowy Program Rozwoju Humanistyki
- ❧ Okres finansowania: 2013–2018
- ❧ Jednostka koordynująca: IJP PAN
- ❧ Kierownik: Włodzimierz Gruszczyński
- ❧ Koordynator: Renata Bronikowska
- ❧ Kryptonim: KorBa (**K**orpus **B**arokowy)
- ❧ Współpraca: IPI PAN (zwłaszcza projekt *Chronofleks* kierowany przez M. Wolińskiego)
- ❧ Objętość: 12 mln segmentów
- ❧ Objętość podkorpusu znakowanego ręcznie: 0,5 mln segmentów



Najważniejsze założenia



- ❧ Bogate metadane:
 - 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
- ❧ Reprezentatywność chronologiczna, geograficzna, genologiczna itp.
- ❧ Oznakowanie morfosyntaktyczne



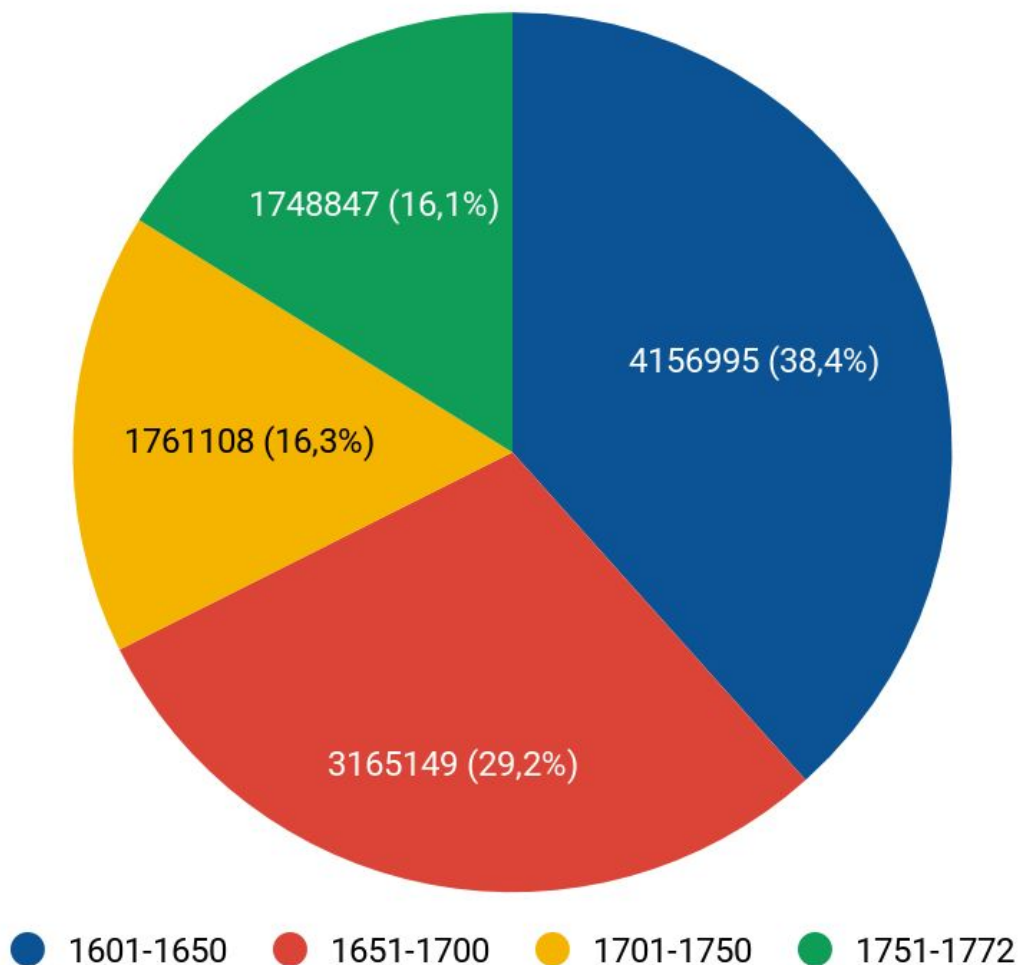
Obecny stan korpusu



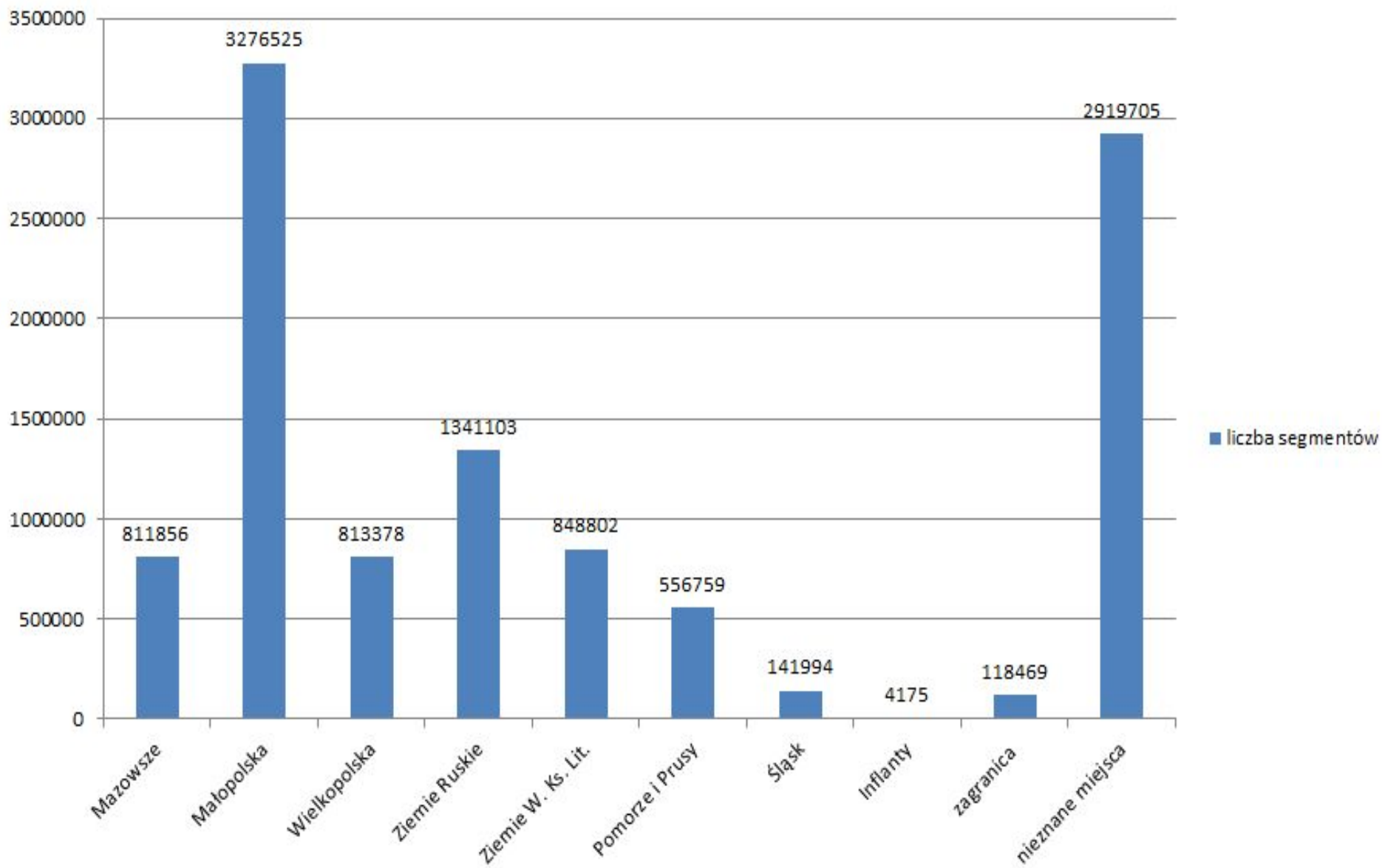
- Wprowadzono 718 tekstów o łącznej długości 10,97 mln słów, co daje ponad 12 mln segmentów w rozumieniu NKJP.
- Wszystkie teksty mają dokładne metryczki.
- Wszystkie teksty mają oznakowaną strukturę i oznaczone słowa w językach obcych.
- Niektóre teksty wymagają ciągle korekty ☹.
- Działają następujące narzędzia:
 - Wczytywacz,
 - Transkryber,
 - Anotatornia 2,
 - Korbeusz, czyli postarzony Morfeusz,
 - Wyszukiwarka MTAS (zmodyfikowana).



Reprezentacja chronologiczna tekstów



Reprezentacja regionalna tekstów



Charakterystyka genologiczna i tematyczna

KorBa

KORBA — wczytywacz tekstów

włodek (Redaktor)

[Strona główna](#)

[Nowy tekst](#)

[Wyszukiwarka](#)

[Raporty](#)

Pracuj jako:

[skryptor](#)

[korektor](#)

[Zgłoś błąd](#)

[Zmień hasło](#)

[Wyloguj](#)

Wczytanych tekstów: ,
zawierających w sumie (wg
TEI) słów.

Zakończonych tekstów: ,
zawierających w sumie (wg
TEI) słów.

typ mowy:

rodzaj:

gatunek:

- ☐ akta sejmikowe ☐ bajki ☐ dedykacje ☐ dialogi ☐ diariusze sejmowe ☐ emblematy ☐ encyklopedie, kompendia
☐ epitafia ☐ epitalamia ☐ fraszki i epigramaty ☐ herbarze ☐ instrukcje ☐ interczy ☐ inwentarze ☐ jasełka
☐ kalendarze ☐ kazania ☐ kolędy i pastorałki ☐ komedie ☐ konstytucje sejmowe ☐ kroniki ☐ księgi liturgiczne
☐ księgi sądowe ☐ lamenty ☐ listy urzędowe ☐ modlitwy ☐ mowy okolicznościowe ☐ mowy polityczne ☐ ody
☐ opisy geograficzne ☐ opisy podróży ☐ pamiętniki ☐ panegiryki ☐ pieśni ☐ pisma polityczne, społeczne
☐ pisma religijne ☐ podręczniki ☐ poematy epickie ☐ poradniki ☐ przepisy kucharskie ☐ przewodniki
☐ przypowieści, specula (zwierciadła) ☐ przysłowia ☐ przywileje, akty nadania ☐ psalmy ☐ relacje ☐ romanse
☐ rozmówki do nauki języka ☐ satyry ☐ sielanki ☐ sonety ☐ teksty sejmowe ☐ testamenty ☐ tragedie
☐ traktaty ☐ umowy ☐ wilkierze ☐ zagadki ☐ żywoty świętych

tematyka:

- ☐ alchemia ☐ anatomia ☐ architektura ☐ astrologia ☐ astronomia ☐ biologia ☐ botanika ☐ budownictwo
☐ chemia ☐ egzotyka ☐ ekonomia ☐ filozofia ☐ fizyka ☐ geografia ☐ gospodarstwo ☐ gramatyka ☐ górnictwo
☐ historia ☐ hutnictwo ☐ języki ☐ kulinaria ☐ magia ☐ matematyka ☐ medycyna ☐ mitologia ☐ muzyka
☐ myślistwo ☐ obyczajowość ☐ poetyka ☐ polityka ☐ prawo ☐ religia ☐ retoryka ☐ sprawy sądowe
☐ wojskowość ☐ zieleństwo ☐ zoologia ☐ zeglarstwo

poetyka żartu: ☐

region:

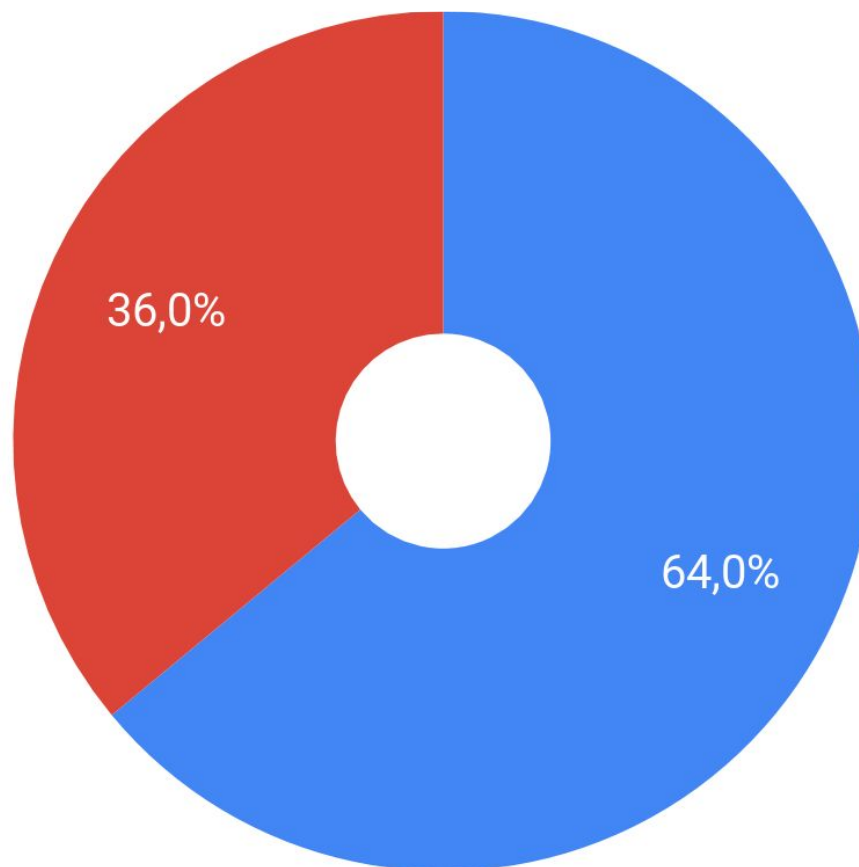
50-lecie:

uwspółcześnienie:

Typy tekstów w korpusie



- Starodruki i rękopisy
- Wydania współczesne



Wyzwania stawiane przez teksty dawne



ce pokoim ozłaca. Roskaza te Niebá Boszkow
wie/ áby Niebá krwáwỹ/ żeláznym dżdżem ná
nieprzyiaćioł lały/ áby z swych cekausow gro
mow dobyły / áby wykretnego Szwedá ogłu
szyły; áby Węgrzyná postrąszyły. Spráwne

ce pokoim ozłaca. Roskaza te^o^ Niebá
Boszkowie/ áby Niebá krwáwỹ/ żeláznym
dżdżem ná nieprzyiaćioł lały/ áby z swych
cekausow gromow dobyły/ áby wykretnego
Szwedá ogłuszyły; áby Węgrzyná
postrąszyły. Spráwne

Wyzwania stawiane przez teksty dawne

Tayneli wam są? rzecze, te Tyraństwa
 Pod ktore Cazar wszystkich nas naginał
 I z miłych swobod pędząc do poddaństwa
 Na Thron dziedziczny bezpiecznie się wspinał,
 Niechcę Authorem znać go mego Państwa,
 Ia bym sam był w nim utopił puginał
 Anibym go był cierpiał w Maiestacie
 Choć mnie postawił on w tym Konsulacie.



Tayneli Wam są? rzecze, te Tyraństwa
 Pod ktore Caezar wszystkich nas naginał
 I z miłych swobod pędząc do poddaństwa
 Na Thron dziedziczny bezpiecznie się wspinał,
 Niechcę Authorem znać go mego Państwa,
 Ia bym sam był w nim utopił puginał
 Anibym go był cierpiał w Maiestacie
 Choć mnie postawił on w tym Konsulacie.

Komentarze i poprawki

- ☞ Informacja o liczbie nieczytelnych znaków
 - COMED<..>
- ☞ Informacja o możliwości błędnego odczytania słowa
 - gospodarz <?>
- ☞ Informacja o zaskakującym zapisie
 - gorzski <!>
- ☞ Rozwinięcie skrótów oznaczonych tyldą
 - atramēt <atrament>
- ☞ Poprawka oczywistej literówki
 - wszzędze <wszędzie>

Znakowanie struktury tekstu i fragmentów obcojęzycznych

- ☛ Oznaczenia elementów struktury tekstu, np.
[POCZĄTEK NOTKI] [KONIEC NOTKI]
- ☛ Oznaczenia fragmentów pominiętych, np.
[FRAGMENT NIECZYTELNY]
- ☛ Oznaczenia pewnych typów fragmentów tekstu, np.
Tytuł części, *Streszczenie części*
- ☛ Oznaczenia języków obcych, np. **łacina**, **niemiecki**

Jaki to język?

**Modlitwa
Tatarska.**

**La illach illechach/ Machemet ire ūl allach/
tanrebir pagamber hach. ix pix.**

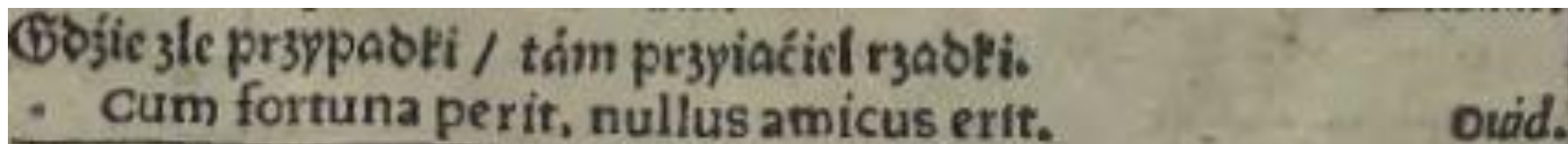
**O xpádnieniu sefty Máchometáństiey y páństwá
Tureckie^oy Tatársk. Proroctwo od sámych Turpów mydáne.
Patissahomoz ghlur/ Ciaferun memleketi alur /
feuzul almai alur/ kapzeyler/ iedi yladeth Gvaur
feleci Csiřmasse/on iři yladegh onlaron beghligh
eder: eufiıapar/ baghi difer / bahciıai baghlar/
ogli fezi olur: onıři yldenřforá Ğristianon feleci
Csiřar/ ol Turki ghereřfine tus fure.**

Leñior meus salve.

**Auctor Les
etori.**

**Liberae sunt voces & cogitationes Polonorum & Lithuanorum. contra
Tartaros, Christiani nominis hostes Vale.**

Od skanu do formatu TEI XML



Gdzie złe przypadki / tam przyjaćiel rzadki.

Cum fortuna perit, nullus amicus erit.

[POCZĄTEK NOTKI] Ouid. [KONIEC NOTKI]

TEI XML



Gdzie złe przypadki / tam przyjaćiel rzadki.

<foreign xml:lang= "la" xml:id= "txt_12.7.7-foreign" >

Cum fortuna perit, nullus amicus erit.</foreign>

<note place= "margin" xml:id= "txt_12.7.8-note" >Ouid.</note>

Klasy fleksyjne – różnice w stosunku do NKJP

	NKJP			KorBa		
Leksem	Fleksem	Ozn.	Przykład	Fleksem	Ozn.	Przykład
Rzeczownik	• rzeczownik • forma depr.	subst depr	<i>królowie</i> <i>króle</i>	Rzeczownik	subst	<i>królowie, wilcy, anioły, anieli</i>
Liczebnik przymiotnikowy				liczebnik przymiotnikowy	adjnum	<i>drugi, dwukrotny, dwojaki, samowtóry</i>
Liczebnik przysłówkowy				liczebnik przysłówkowy	advnum	<i>dwukrotnie, dwojako, dwakroć, samowtór</i>
Przymiotnik	• przymiotnik • przymiotnik poprzyminkowy • przymiotnik przyprzymiot. • przymiotnik predykatywny	adj adjp adja adjc	<i>polski</i> <i>polsku</i> <i>polsko</i> <i>wesół</i>	• przymiotnik • przymiotnik odm. niezłożona • przymiotnik przyprzymiot.	adj adjb adja	<i>dobry, zdrowego;</i> <i>zdrów, gwałtownę,</i> <i>polsku;</i> <i>angielsko, ziemno</i>
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	<i>był</i> <i>będzie</i> <i>-(e)ch, -(e)chmy</i> <i>będący, jadący</i> <i>umęczon,</i> <i>ukrzyżowan</i>

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, pokole-niu C.N. (dwie) mężoma, zakonoma; (dwie) niewiastama, rze-czoma; (dwie ma) 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 być następujący:

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

Rodzaj

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

Kategorie gramatyczne: rodzaj męski żywotny

- ☞ Wiele rzeczowników rodzaju mianim może mieć dwie konkurencyjne formy pl:nom, np.:
 - (te) *lwy, ptaki, anioły, pany, chłopcy,*
 - (ci) *lwowie, ptacy, anieli||aniołowie, panowie, chłopci.*
- ☞ Formom tym przypisujemy odpowied-nio rodzaj m i mianim1.
- ☞ 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 mianim1, np. z formami typu *ci, wielcy, byli.*

Przygotowanie próbek do ręcznej anotacji

Proces przygotowania próbek korpusowych obejmuje:

- ☞ podział tekstu na zdania,
- ☞ transkrypcję,
- ☞ analizę morfosyntaktyczną,
- ☞ zapis do plików zgodnych z TEI XML.

Powyższe etapy wykonywane są automatycznie i sterowane przez jeden program (tzw. **Amebę**).

Następnie próbki trafiają do systemu **Anotatornia 2** w celu ręcznego ujednoznacznienia.

Automatyczny podział próbki na zdania

Do podziału na zdania wykorzystano program **srx_segmenter** (https://github.com/narusemotoki/srx_segmenter).

Pozwala on zdefiniować za pomocą wyrażeń regularnych konteksty, w których koniec zdania ma nastąpić (np. kropka, po której jest spacja i wielka litera) oraz wyjątki od tych reguł (np. skróty, inicjały pisane z kropką, które nie mogą kończyć zdania itp.).

Dzięki tym regułom poprawnie analizują się np. zdania:

- ☞ S. Suplikuię WX. Mci ażebyś mi kazał uczynić sprawiedliwość.
- ☞ Iest też tám zgromádne Collegium Conuictorow/ kędy iest miedzy inszymi około 40. Zakonnikow z rozmáitych zakonow/ á osobliwie ś. Benediktá.

Automatyczna transkrypcja

Transkrypcja oparta jest na regułach zapisanych wyrażeniami regularnymi (wykorzystano <https://bitbucket.org/jsbien/pol>).

Kont. lewy	Do zamiany	Kont. prawy	Po zamianie	Przykład przed	Przykład po
.*	é	.*	e	potém, którém	potem, które
.*	th	.*	t	theatrum, Lutherani	teatrum, Luterani
^	rown	.*	równ	rownego, równość	równego, równość
^	iako	\$	jako	iako	jako
A	y	\$	j	bardziej, zwyczaj	bardziej, zwyczaj

.* – dowolny ciąg znaków (również pusty)

^ – początek słowa

\$ – koniec słowa

A – dowolna samogłoska (zbiór zdefiniowany w szablonie)

Kiedy reguły zawodzą

- ☛ Jeśli reguła „łapie” zbyt dużo, można wprowadzić wyjątki.

Kont. lewy	Do zamiany	Kont. prawy	Po zamianie	Przykład przed	Przykład po
A	i	A	j	wzywaią, nadzieię, iii	wzywają, nadzieję, iji

Wykluczenie z tej reguły liczb rzymskich (jedyną samogłoską jest i): **i+**

- ☛ Jeśli trudno zapisać ogólną regułę (nawet z wyjątkami), pozostaje wyliczenie i zamienianie całych słów.

Kont. lewy	Do zamiany	Kont. prawy	Po zamianie	Przykład przed	Przykład po
^	poki	\$	póki	poki	póki
^	zeby	\$	żeby	zeby	żeby

Kontrola i poprawianie transkrypcji

Transkrypcja wykonana automatycznie sprawdzana jest w czasie ręcznej anotacji. Poprawki mogą być konieczne z powodu:

- ☛ Błędów lub braków w regułach transkrypcji (w trakcie pracy nad anotacją wielokrotnie reguły były aktualizowane)
- ☛ Pomyłek w czasie przepisywania lub nietypowego zapisu w oryginale
- ☛ Sytuacji, w których transkrypcja zależy od kontekstu całego słowa (np. słowo **v** może pozostać w niezmienionej formie jako liczba rzymska lub być przetranskrybowane na **u**).

W korpusie znakowanym ręcznie anotatorzy ingerowali w transkrypcję wykonaną automatycznie w **2,4%** przypadków.

Analiza morfologiczna

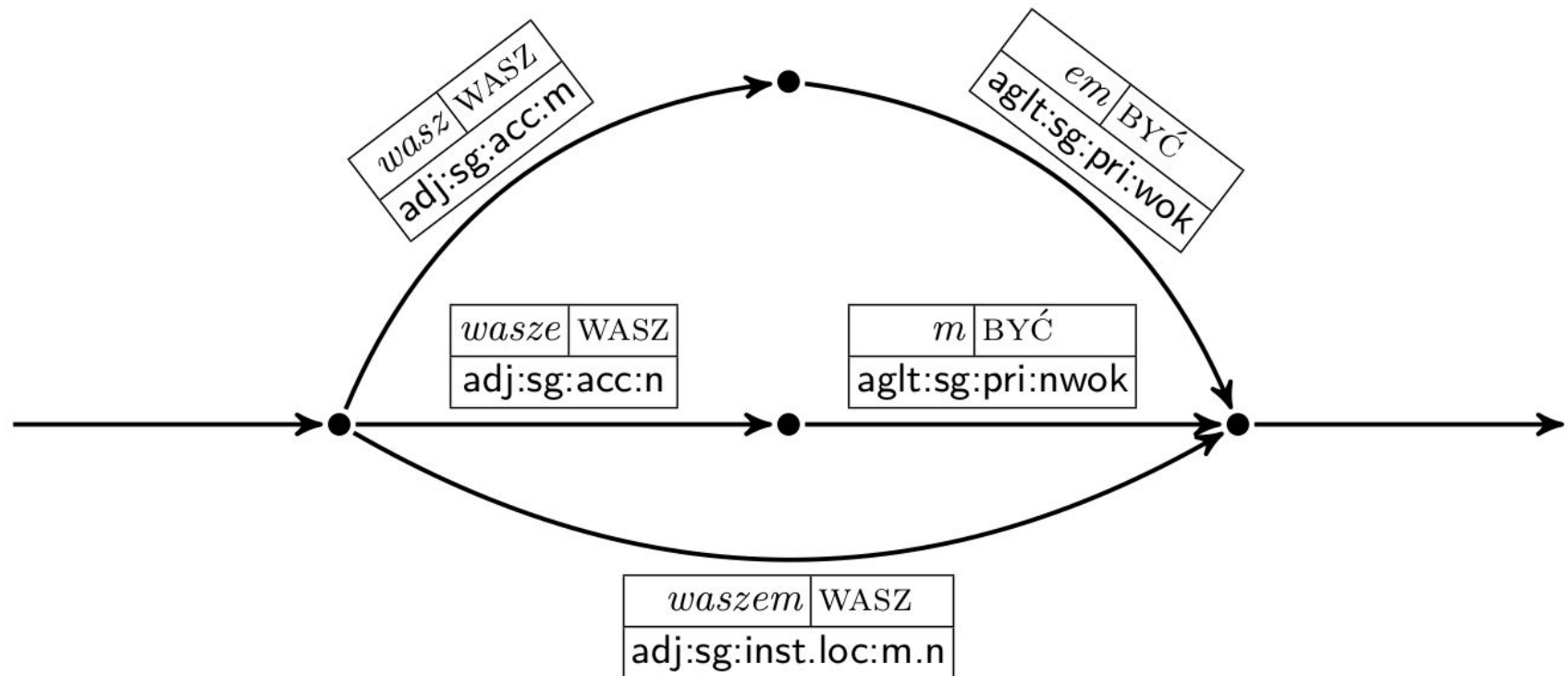
Słownik:

- ☞ „postarzone” dane SGJP — konwersja tagsetu i dotworzenie niektórych regularnych dawnych form (*walidaczej, ładnę, ładnemi, zmienim, klęczta, klęczwa* itp.),
- ☞ dane fleksyjne SXVII,
- ☞ automatycznie zrekonstruowane częściowe paradygmaty SXVII.

Reguły segmentacyjne:

- ☞ pisownia łączna z *nie*,
- ☞ pisownia łączna licznych partykuł (*ć, li, że, ż*),
- ☞ znacznie bardziej ruchome aglutynanty.

Niejednoznaczna segmentacja



Anotatornia 2 – ręczna anotacja

System **Anotatornia 2** umożliwia:

- ☞ import próbek w formacie TEI XML z wprowadzonymi informacjami morfosyntaktycznymi, podziałem na zdania i transkrypcją,
- ☞ zmianę podziału na zdania,
- ☞ zmianę transkrypcji segmentów,
- ☞ wybór właściwego wariantu segmentacyjnego,
- ☞ łączenie lub rozdzielanie segmentów,
- ☞ wybór jednej z zaproponowanych interpretacji lub wpisanie własnej,
- ☞ pracę w systemie **2 anotatorów + superanotator** lub **pojedynczego anotatora**,
- ☞ eksport próbek w formacie TEI XML z oznaczeniem wybranej interpretacji.

Anotatornia 2 – widok anotatora

DorHip_II, próbka 2

Uwagi Gotowe

Tekst transliterowany

Podkowę| iedną| w| tym| przypadku| pamiętay| odiać|/| iżeby| stał| miękko|/| ále| ná| **suszy**|.↵

NA| zátretowánie|/| áby| się| potym| kopyto| nie| záropiło|/| skoro| ono| postrzeżesz|/| zarázém| iáko| nalepiey| krew| wyćiśni| ręką|/| á| octem| z| solą| przemyi|/| á| ná| noc| grzanke| ze| rżanego| chlebá| vpiekszy| przyłoż|:| názáiutrz| zámáć| pszenicznej| mąki| z| białkiem| iáiwym|/| á| przyłożywszy|/| do| trzeciego| dnia| zániechay|.↵

Weźmi| owczych| bobkow|/| oleiu| lniánego|/| octu| twárdego|/| Máiwego| máslá|/| vtłukszy| bábki| rospuść| w| tym| dobrze|/| iż| będzie| iáko| máść|/| rozmaż|ze| iá| ná| chustę|/| á| tym| nogę| y| zátret| obwiń|.↵

Potym| vwoszczywszy| dobrze| płotno|/| záwiń| po| wierzchu|/| żeby| wodá| nie| doszłá| zwłaszczá| iesli|ć| w| drogę| iecháć| przydźie|.↵

Weźmi| ciepłego| chlebá| z| piechá| wyiętego|/| á| zmieszawszy| z| solą|/| przywiń| co| dzień| świeżego|/| áż| się| zgoi|.↵

Inszy| też| tylko| błotem| okładáia| co| raz| ponawiáiać| gdy| oschnie|.↵

Pod| rána| wykárbowawszy| nożem| kowálskim|/| žeby| ropá| schodźiła| ná| do|/| wymyi| ránę| octem| ciepłym| osolonym|:| á| zbiwszy| białku| z| hálunem|/| przyłoż| przez| noc|/| názáiutrz| zaskwarz| łoiem| kozłowym|/| á| muśisz|li| w| drogę| iácháć|/| smołą| czarną| po| wierzchu| zásmol| ciepło|/| by| wodá| nie| doszłá|.↵

suszy

susza subst sg:loc:f

Lemat: susza

Cz.m.: subst

Tag: sg:loc:

OK

sg:loc:m

sg:loc:f

sg:loc:n

☐ susz subst pl:gen:m

☐ susza subst sg:gen:f

☐ susza subst sg:dat:f

☒ susza subst sg:loc:f

☐ susza subst pl:gen:f

☐ suszyć fin sg:ter:imperf

.

interp

Anotatornia 2 – widok superanotatora

DorHip_II, próbka 2

Gotowe ✓

Tekst transliterowany

Podkowę| iednák| w| tym| przypadku| pámiętay| odiáč|/| **iżeby**| iże|by| iże|by|
stał| miękkó|/| ále| ná| suszy|.↵

NA| zátretowáníe|/| áby| się| potym| kopyto| nie| záropíto|/| skoro| ono|
postrzeżesz|/| **zárázem**| iáko| nalepiey| krew| wyčíśni| ręká|/| á| octem| z| solá|
przemyi|/| á| ná| noc| grzankę| ze| rżanego| chlebá| vpiekszy| przylož:| názáiutrz|
zámać| pszeniczney| mąki| z| białkiem| iáíowym|/| á| przyložyszy|/| do| trzeciego|
dniá| zániechay|.↵

Weźmi| owczych| bobkow|/| oleiu| Iníanego|/| octu| twárdego|/| Máíowego|
máslá|/| vtłukszy| **bábki**| rospuść| w| tym| dobrze|/| iž| bédźie| iáko| máśc|/|
rozmaž|že| iá| ná| chustę|/| á| tym| nogę| y| zátret| obwiń|.↵

Potym| wvoszczywszy| dobrze| płotno|/| záwiń| po| wierzchu|/| **žeby**| wodá| nie|
doszłá| **zwłaszczá**| iesli|ć| w| drogę| iechác| przyjdzie|.↵

Weźmi| ciepłego| chlebá| z| piechá| wyiętego|/| á| zmieszawszy| z| solá|/| przywiń|
co| dzień| świeżego|/| áž| się| zgoi|.↵

Inszy| też| tylko| błotem| okładáią| **co raz**| ponawiáiąc| gdy| oschnie|.↵

Pod| ráną| wykárbowawszy| **nożem**| kowálskim|/| **žeby**| ropá| schodžítá| ná| do|/|
wymyi| ránę| octem| ciepłym| osolonym|:| á| zbiwszy| **białku**| **z**| háłunem|/| przylož|
przez| noc|/| názáiutrz| zászwarz| łoiem| kozłowym|/| á| muśisz|li| w| drogę|
iáchác|/| smołá| czarná| po| wierzchu| zásmol| ciepłó|/| by| wodá| nie| doszłá|.↵

/

/ interp

iżeby

iżeby comp

iże

iże comp

by

by qub

iże

☐ iże comp

by

☐ by qub

stał

stać praet sg:m:imperf

miękkó

miękkó adv pos

Tagowanie

☛ Concraft:

- ☛ conditional random fields,
- ☛ ujednoznacznianie niejednoznacznej segmentacji (~94%),
- ☛ słabsze wyniki dezambiguacji (~88%),
- ☛ nieco lepsze zgadywanie.

☛ Toygger:

- ☛ sieci neuronowe (bi-LSTM),
- ☛ wyłącznie ujednoznacznianie znaczników na zadanej segmentacji,
- ☛ lepsze wyniki dezambiguacji (~91%),
- ☛ nieco słabsze zgadywanie.

Elektroniczny Korpus Tekstów Polskich XVII i XVIII w. – perspektywy



- ❧ Złożony został wniosek o sfinansowanie kontynuacji projektu przez Narodowy Programu Rozwoju Humanistyki.
- ❧ Okres finansowania: 2019-2024.
- ❧ Jednostka koordynująca: IJP PAN.
- ❧ Kierownik: Włodzimierz Gruszczyński.
- ❧ Koordynator: Renata Bronikowska.
- ❧ Kryptonim: KorBa2 .
- ❧ Współpraca: IPI PAN.
- ❧ Powiększenie objętości o 12 mln segmentów
- ❧ Wprowadzenie nowych tekstów i uzupełnienie tekstów obecnie dostępnych we fragmentach.
- ❧ Uwzględnienie ostatniego ćwierćwiecza XVIII w.
- ❧ Eksperymenty m.in. ze znakowaniem składniowym i semantycznym.



KorBa

*Teraz przy zakończeniu tey przemowy wszyscy
authorowie, uniżony pokłon oddając,
J.W.W.Mćiom Pańiom y Panom za ich
wysłuchanie z gorącą attencyą pokornie dziękuią
y życzą J.W.W.Mćiom zupełney obfitości
wselákiego szczęścia.*



W Wárśáwie, dnia 23 Aprila A.D. CIOCIØXVIII