Tagset barokowy

– problemy opracowania zestawu kategorii morfologicznych i ich wartości na potrzeby Elektronicznego Korpusu Tekstów Polskich XVII i XVIII w. (do 1772 r.)

Włodzimierz Gruszczyński
Instytut Języka Polskiego PAN
wlodekiewa@poczta.onet.pl

KorBa

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

- Finansowanie projektu: Narodowy Program Rozwoju Humanistyki na lata: 2013-2018.
- Jednostka koordynująca: IJP PAN.
- Kierownik: Włodzimierz Gruszczyński.
- Koordynator: Renata Bronikowska.
- Kryptonim: KORBA (**KOR**pus **BA**rokowy).
  - Współpraca: IPI PAN (projekt ChronoFlex, kierowany przez M. Wolińskiego).
  - Objętość: 12 mln segmentów.
  - Objętość podkorpusu znakowanego ręcznie: 0,5 mln segmentów.

Obecny stan korpusu

- Wprowadzono ponad 717 tekstów.
- Łączna długość tekstów 10,94M segmentów, co daje ponad 12M w rozumieniu NKJP.
- Wszystkie teksty mają dokładne metryczki.
- Wszystkie teksty są oznakowane strukturalnie i językowo.
- Niektóre teksty wymagają korekty, którą obecnie prowadzimy.
- Działają następujące narzędzia:
  - Wczytywacz
  - Transkryber (uzupełniony został zestaw reguł).
  - Anotatornia2.
  - Morfeusz'17 (systematycznie modyfikowany).
- Wersja testowa dostępna pod adresem: http://korba.edu.pl/.

Najważniejsze założenia – bogate metadane

- Dokładne dane bibliograficzne wszystkich tekstów.
- Charakterystyka stylistyczno-genologiczna.
- Charakterystyka tematyczna tekstów.
- Charakterystyka socjolingwistyczna i geolingwistyczna (informacje o autorach, wydawcach, tłumaczach).

Najważniejsze założenia – znakowanie wszystkich tekstów

- Znakowanie strukturalne umożliwiające dokładną lokalizację cytatu w tekście:
  - paginacja (uwzględnienie różnych systemów),
  - oznaczenia granic rozdziałów, części, ksiąg itp.,
  - oznaczanie notek marginesowych, przypisów itp.,
  - oznaczanie podpisów pod ilustracjami;
  - oznaczanie typowych części, np. dedykacyj, typowych fragmentów strony tytułowej itp.
  - oznaczanie opuszczeń zapisów nutowych, rysunków, wzorów, symboli.

- Znakowanie językowe, tzn. oznaczenie wszystkich wtrętów obcych, z podziałem na języki.

Cele ręcznej anotacji morfologicznej

- Utworzenie niewielkiego względnie zrównoważonego korpusu, w którym segmenty są rozpoznane przez kompetentne osoby i oznakowane konsekwentnie według tej samej instrukcji.
- Utworzenie tzw. korpusu treningowego dla automatycznego tagera, za pomoca którego oznakowany zostanie korpus 12-milionowy.
  - Cel drugi jest celem zasadniczym!
  - Tager automatyczny działa na podstawie danych statystycznych, więc kategorie nie mogą być zbyt szczegółowe i trudne do rozpoznania.

Podkorpus przeznaczony do znakowania ręcznego

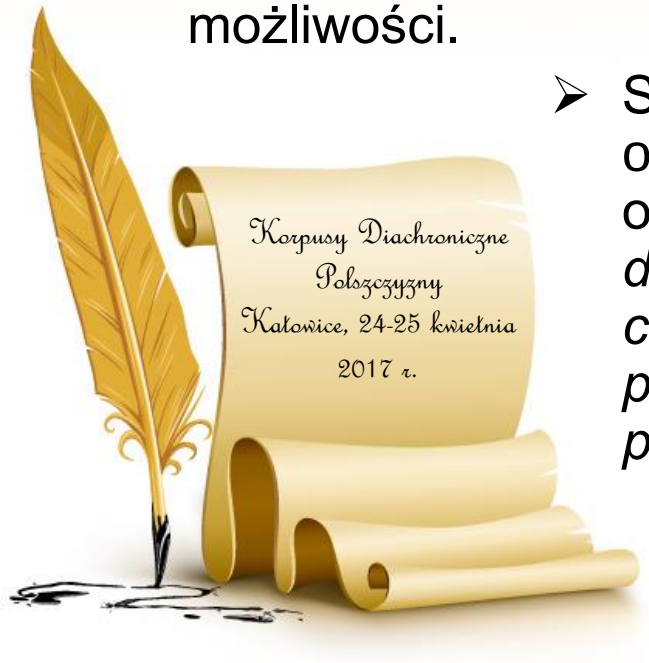
- Planowana objętość: 0,5 mln segmentów.
- Podkorpus będzie reprezentatywny dla korpusu głównego.
- Próbki losowane są ze wszystkich tekstów, przy czym liczba próbek z jednego tekstu jest proporcjonalna do jego długości.
- Próbki mają długość ok. 200 segmentów, zaczynają się i kończą pełnymi zdaniami.

Główne założenia dotyczące znakowania i segmentacji

- Znakowanie morfologiczne (morfosyntaktyczne) ma być możliwie zbliżone do znakowania w Narodowym Korpusie Języka Polskiego.
- Punktem wyjścia instrukcji znakowania była instrukcja do znakowania w NKJP.
- Konieczne było uwzględnienie kategorii i wartości kategorii gramatycznych nieistniejących we współczesnej polszczyźnie.
- Konieczne jest uwzględnienie tego, że anotatorzy nie są native-speakerami XVII-wiecznej polszczyzny, więc nie mogą stosować testów, w których rozstrzygająca jest ocena poprawności – niepoprawności frazy.

Segmentacja w próbkach

- Segmentacja tekstu przebiega na dwóch poziomach: podział na zdania (równoważniki) i na jednostki wyrazopodobne, tzw. segmenty.
- Segmentacja na zdania wykonywana jest automatycznie, ale anotatorzy mogą ją zmieniać.
- Segmentacja na segmenty dokonywana jest również automatycznie, anotatorzy mogą ją zmieniać, a w pewnych wypadkach muszą dokonać wyboru z dwóch dopuszczanych przez Anotatornię możliwości.



- Segmenty nie zawsze odpowiadają słowom ortograficznym (podobnie jak w NKJP), np.: osobnymi segmentami są człony słów *łgał*|*eś*, *to*|*ś*, *długo*|*śmy*, *tak*|*em*, *jakoby*|*śwa*, *gdy*|*śwa*, *już*|*eśta*, *coś*|*ta*; *przyszedł*|*by*, *napisała*|*by*|*m*, *chodź*|*że*, *potrzebował*|*że*|*by*|*ś*, *(nie) kwapcie*|*ż (się)*, *znasz*|*li*, *pragnę*|*ć.*

Podstawowe informacje na temat znakowania

- Każdemu segmentowi tekstu przypisywany jest przez analizator morfologiczny Korbeusz znacznik (lub, w wypadku segmentów o kilku możliwych interpretacjach, kilka znaczników) interpretujący dany segment jako wykładnik tekstowy pewnej formy wyrazowej.
- Znaczniki zawierają dwa rodzaje informacji:
  - formę podstawową (tzw. lemat),
  - charakterystykę morfoskładniową danego segmentu (tzw. znacznik morfosyntaktyczny).

- Zadaniem anotatora jest wykonanie jednej z trzech czynności:
  - potwierdzenie adekwatności „podpowiedzianego” znacznika,
  - wybór jednego z wielu możliwych znaczników,
  - wpisanie samodzielnie utworzonego znacznika.

Struktura znaczników morfosyntaktycznych

- Znacznik morfosyntaktyczny jest ciągiem wartości rozdzielonych dwukropkami, np.: `subst:sg:acc:f` dla segmentu *bańkę*.
  - Pierwsza wartość określa **klasę fleksyjną** (`subst:` rzeczownik),
  - Następne – wartości **kategorii gramatycznych** przysługujących danej formie (`sg:acc:f`).

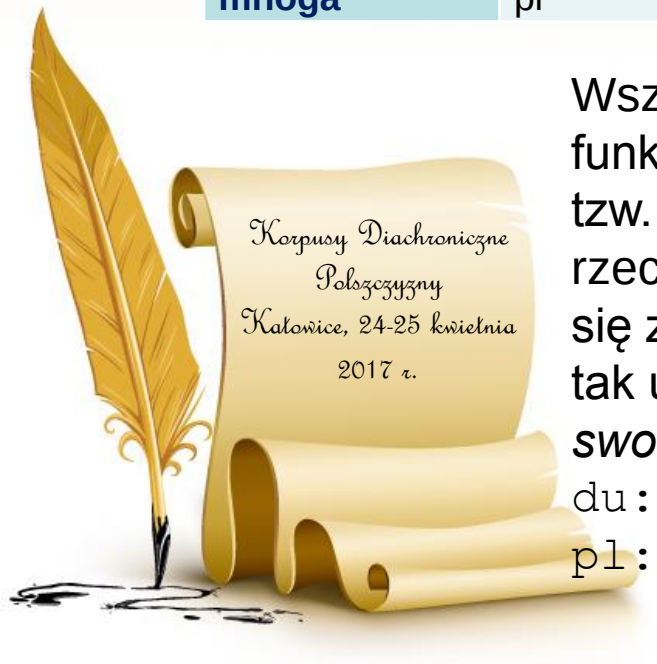
- W wypadku, gdy dla danego segmentu występuje niejednoznaczność wartości jakiejś kategorii, podawanych jest kilka znaczników, np.: `subst:sg:dat:f` i `subst:sg:loc:f` dla segmentu *osobie*.

Klasy fleksyjne – różnice w stosunku do NKJP

leksem	fleksem	ozn.	przykład
rzeczownik	rzeczownik	subst	woda,
liczebnik	liczebnik główny	num	pięć, jedenaście
	liczebnik zbiorowy	numcol	pięcioro, dwojga
liczebnik przymiotnikowy	liczebnik przymiotnikowy (porządkowy, wielokrotny, mnożny, wieloraki)	adjnum	drugi, dwukrotny, podwójny, dwojaki, samowtóry
liczebnik przysłówkowy	liczebnik przysłówkowy (wielokrotny, mnożny, wieloraki)	advnum	dwukrotnie, podwójnie, dwojako, dwakroć, samowtór
przymiotnik	przymiotnik	adj	dobry,
	przymiotnik odm. niezłożona	adjb	gwałtownę, polsku, zdrów
	przymiotnik przyprzymiotnikowy	adja	angielsko, ziemno
przysłówek		adv	bardziej, kiedy
zaimek	nietrzecioosobowy	ppron12	ja, ty, my, wy
	trzecioosobowy	ppron3	on, ona, ono, oni, one
	SIEBIE	siebie	siebie, sobą, sobie
czasownik	forma nieprzeszła	fin	czytam
	forma przyszła czasownika BYĆ	bedzie	będę, będziesz (w domu)
	forma przeszła czasownika BYĆ (składowa czasu zaprzeszłego)	plusq	był
	forma czasownika być jako składnik czasu przyszłego	fut	będę (pisał, pisać)
	aglutynant czasownika BYĆ	aglt	-śmy
	aglutynant „aorystyczny” cz. BYĆ	agltaor	-(e)ch, -(e)chmy
	pseudoimiesłów	praet	czytał
	rozkaźnik	impt	czytaj
	bezosobnik	imps	czytano
	bezokolicznik	inf	czytać
	imiesłów przysłówkowy współczesny	pcon	czytając
	imiesłów przysłówkowy uprzedni	pant	zjad(ł)szy, widziawszy
	odsłownik	ger	czytanie
	imiesłów przymiotnikowy czynny	pact	czytający, bolątszy
	imiesłów przymiotnikowy czynny odm. niezłożona	pactb	będęcy, jadęcy
	imiesłów przymiotnikowy bierny	ppas	czytany
	imiesłów przymiotnikowy bierny odm. niezłożona	ppasb	umęczon, ukrzyżowan
	imiesłów przeszły	ppraet	osiwiały, były
czasownik typu WINIEN (forma teraźniejsza)		winien	winna, powinni
predykatyw		pred	trzeba

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 bez względu na ich funkcję i składnię znakujemy jako du. Chodzi np. o związek z tzw. przydawką przymiotnikową – nawet jeśli forma rzeczownika t końcówce typowej dla liczby podwójnej łączy się z przymiotnikiem w liczbie mnogiej (a nie podwójnej), to i tak uznajemy ją za formę dualis, np. we frazie (*robił to*) *swoimi starymi rękoma* formę *rękoma* znakujemy jako `du:instr`, chociaż przymiotniki znakujemy jako `pl:instr:f:pos`.

Kategorie gramatyczne: rodzaj

- W ustalaniu rodzaju rzeczowników w języku średniopolskim nie mogą pomóc konteksty diagnostyczne podobne do tych, 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), dlatego w wypadkach segmentów reprezentujących nieznane anotatorowi leksemy lub nieznane w formie występującej w próbce należy przypisywać rodzaj „uogólniony”. Tryb przypisywania rodzaju powinien być następujący:
  - formom leksemów znanych anotatorowi i zgodnym ze współczesną normą – wg jego intuicji;
  - formom znanym, do których są różne podpowiedzi (niezgodne ze stanem wsp.) – wybieramy wartość wspólną (możliwą do uzgodnienia z każdą ewentualnością), 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 są podpowiedzi różniące się rodzajem, przypisujemy rodzaj wspólny;
  - 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 nosob.	p2	(grać w) arcaby (lemat ARCABY) (podał) abdyta (lemat ABDYTA)

Kategorie gramatyczne rodzaj męski żywotny

- 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*.

- Identycznie traktujemy synkretyczne z nimi formy wołacza l.mn.
- Rodzaj `manim1` przypisujemy poza tym tylko formom `pl:acc` identycznym z dopełniaczem (np. *widzę tych konsulów*).

Kategorie gramatyczne
aspekt

Dziękuję za uwagę!