

NAKŁADKA KORPUSOWA

OPARTA O TRADYCYJNĄ KLASYFIKACJĘ CZĘŚCI MOWY
(NKJP, KORBA)

Emanuel Modrzejewski
modrzejewski.emanuel@gmail.com

DOTYCHCZASOWE NAKŁADKI KORPUSOWE:

- Polsko-rosyjski i rosyjsko-polski korpus równoległy
- Korpus języka młodzieży początku XXI wieku



 Forma słownikowa:

▼ Części mowy

☐ Pokaż opcje

☒ Rzeczownik

☐ Czasownik

☐ Przymiotnik

☐ Liczebnik

☐ Spójnik

☐ Przysłówek

☐ Zaimek

☐ Przyimek

☐ Inne

 Forma słownikowa:

▼ Części mowy

☐ Rzeczownik

☐ Przymiotnik

☐ Liczebnik główny

☐ Liczebnik porządkowy

☐ Czasownik

☐ Przysłówek

☐ Predykatyw

☐ Zaimek przymiotnikowy

☐ Zaimek-predykatyw

☐ Spójnik

☐ Partykuła

☐ Wykrzyknik

☐ Część słowa złożonego

Kreator zapytania

Pytanie o segmenty

Pytanie o META dane

Zapytanie statystyczne

[

część mowy

=

rzeczownik (ogólnie)

i

(

przypadek

=

dopełniacz

lub

przypadek

=

biernik

)

i

liczba

=

mnoża

]

[

cokolwiek

]

min: 0

max: 4

[

część mowy

=

czasownik

]

forma hasłowa

forma tekstowa

część mowy

klasa gramatyczna

cokolwiek

liczba

przypadek

rodzaj

Lub

Nie

(

)

Krotność

Usuń

Wyczyść

OK

WADY:

- przełożenie fleksemów 1:1

„ułatwienie” jedynie dla zaawansowanych użytkowników

- wymagana dobra znajomość tagsetu

akomodacyjność, aglutynant, kublik, ...

- brak możliwości wyszukiwania kilku elementów jednocześnie

problem z MWE, elementami nieciągłymi, konkordancjami

- brak obostrzeń = zirytowany użytkownik

czasownik w mianowniku da 0 wyników

ZAŁOŻENIA:

- ułatwienie przeszukiwania, szczególnie dla nielingwistów
np. nazwy własne dla historyków, literaturoznawców, itd.
- stosunkowo łatwa modyfikowalność nakładki w Pythonie i .jsonie
dostosowywanie do różnic między P1 a P2, tworzenie wersji dla korpusów opartych o Poliqarpa
- niwelowanie różnic w tagsetach
możliwość stworzenia wersji działającej na dwóch korpusach
- „intuicyjne” rozumienie części mowy vs. fleksemy
próba wydobywania czasownikowych jednostek nieciągłych np. kategorii gramatycznych typu „czas”,
możliwość tworzenia kwerend opartych o części zdania (prawie)

PROBLEMY:

- kategorie nieprzewidziane w NKJP: czas, tryb, strona
- nieciągłość jednostek

z niedowierzaniem: - Co <u>ście</u> [być:aglt:pl:pri:imperf:nwok] <u>z</u> [z:prep:gen:nwok] <u>siebie</u> [siebie:siebie:gen] <u>zrobiły</u> [zrobić:praet:pl:f:perf]	! Zupełnie nie mógł oswoić
brata i siostrę, żeby <u>m</u> [być:aglt:sg:pri:imperf:nwok] <u>mógł</u> [móc:praet:sg:m1:imperf:nagl]	się z nimi w coś
prezes PZU. - Gdyby <u>śmy</u> [być:aglt:pl:pri:imperf:nwok] <u>mieli</u> [mieć:praet:pl:m1:imperf]	dbać o interes każdego mniejszościowego
No i co niby by <u>m</u> [być:aglt:sg:pri:imperf:nwok] <u>tam</u> [tam:adv] <u>ujrzał</u> [ujrzeć:praet:sg:m1:perf]	? - Gdzie? -
byłby możliwy, gdyby <u>śmy</u> [być:aglt:pl:pri:imperf:nwok] <u>wcześniej</u> [wcześnie:adv:com] <u>nie</u> [nie:qub] <u>zbudowali</u> [zbudować:praet:pl:m1:perf]	instalacji kanalizacyjnej i nie wymienili
mnie spadła gaśnica. My <u>śmy</u> [być:aglt:pl:pri:imperf:nwok] <u>mieli</u> [mieć:praet:pl:m1:imperf]	pałki, górnicy łańcuchy,
Teraz nadszedł czas, by <u>śmy</u> [być:aglt:pl:pri:imperf:nwok] <u>się</u> [się:qub] <u>moralnie</u> [moralnie:adv:pos] <u>zrewanżowali</u> [zrewanżować:praet:pl:m1:perf]	. W Miedźnej wiedzą,

- trudności z transformacyjnością
robił -> był robiony ; (praet: ROBIĆ -> praet: BYĆ + ppas: ROBIĆ)
- bardzo rozbudowane kwerendy
- brak podłączenia do NKJP, mała weryfikowalność

- analityczne stopniowanie przymiotników i przysłówków
bardziej biały | bielszy
- czasy złożone, strona bierna, tryby
niech [on] będzie [dobrze] zrobiony
- osoba
które[ś zrobił] | które zrobił [!agł]
- formy finitywne vs. konstrukcje czasownikowe
był | był zrobiony
- wartości kategorii dla części elementów
będzie robić | będzie robił
- wyrażenia przyimkowe, skróty, burkinostki, partykuły...

RÓŻNICE W TAGSETACH (NKJP vs. KORBA)

- liczebniki przymiotnikowe (adjnum) i przysłówkowe (advnum)

drugi, dwukrotny i podwójnie, dwakroć

- przymiotniki niezłożone (adjb) i przyprzymiotnikowe (adja)

polsku, zdrow i polsko-

- BYĆ w czasie zaprzeszczytym (plusq)

był (zrobił)

- aglutynant „aorystyczny” (agltor)

(by)ch (zrobił)

- imięstów przymiotnikowy czynny (pactb) i bierny (ppasb) w odmianie niezłożonej

będęcy i umęczon

- imięstów przeszły (ppraet)

osiwiały

- fragment, dawna ~burkinostka (frag)

Ali (Pasza)

- liczby arabskie (dig) i rzymskie (romandig)

- liczba podwójna (du)

dwiema niewiastama

- rodzaje:

- nieznany

narańczagami

- męski nieżywotny (mnanim)

stół

- męski żywotny 1 i 2 (manim1, manim2)

tygrysowie i heretyki

- przymnogi 1 i 2 (p1, p2)

państwo i arcaby

- aspekt nieokreślony (biasp)

aresztować

PROBLEMY Z JĘZYKIEM BAROKOWYM:

- z mojego punktu widzenia:
 - interpunkcja (niedzielące kropki w zdaniu, średniki po skrótach)
 - „kreatywność” autorów
 - wtrącenia z języków obcych
 - dystrybucja aglutynantów i partykuł

Nie/ z/tego/ś/ ty/ Pani/ grona/

nie/mógł/że/m/ ja/ in/ omni/ foro/ incorrupto/ tej/ speciem/ fraudis/ dochodzić/
- z punktu widzenia użytkownika:
 - nietypowe realizacje (również dialekty)
 - nieujednolicony zapis

pójdźcie, pódźcie

 - skomplikowany język zapytań (Poliqarp2)

STRUKTURA:

```
},
"branch_verb_cond_pass1_per12" : {
  "neg_cond_pass" : "",
  "aux_past_cond_pass" : "+neg_cond_pass",
  "cond_cond_pass" : "+aux_past_cond_pass",
  "aglt_cond_pass" : "+cond_cond_pass",
  "space_cond_pass" : "+aglt_cond_pass",
  "neg2_cond_pass" : "+space_cond_pass",
  "ppas_cond_pass" : "+neg2_cond_pass"
},
"branch_verb_cond_pass2_per12" : {
  "neg_cond_pass" : "",
  "ppas_cond_pass" : "+neg_cond_pass",
  "space_cond_pass" : "+ppas_cond_pass",
  "neg2_cond_pass" : "+space_cond_pass",
  "aux_past_cond_pass" : "+neg2_cond_pass",
  "cond_cond_pass" : "+aux_past_cond_pass",
  "aglt_cond_pass" : "+cond_cond_pass"
}
},
"specs" : [
  {
    "id" : "neg_cond_pass",
    "quant" : "*",
    "append" : [
      "base=nie"
    ]
  },
  {
    "id" : "neg2_cond_pass",
    "quant" : "*",
    "append" : [
      "base=nie"
    ]
  },
  {
    "id" : "space_cond_pass",
    "quant" : "{,5}",
    "append" : [
      "pos!=\\\"adj||conj||comp||interp||subst||pred\\\"",
      "pos!=verb"
    ]
  }
]
```

budowa gałęzi: wszystkie możliwe kombinacje spełniające założenie, tu: tryb przypuszczający w stronie biernej

specyfikacja poszczególnych tokenów

$[(base=nie)^* [(pos=praet \& base=być) [(base=by \& pos=qub) [(pos!=\\\"adj|conj|comp|interp|subst|pred\" \& pos!=verb)]\{,5\} [base=nie)^* [pos=ppas] | [(base=nie)^* [pos=ppas] [(pos!=\\\"adj|conj|comp|interp|subst|pred\" \& pos!=verb)]\{,5\} [base=nie)^* [(pos=praet \& base=być) [(base=by \& pos=qub)]] | [(base=nie)^* [(pos=praet \& base=być) [(base=by \& pos=qub) [pos=aglt] [(pos!=\\\"adj|conj|comp|interp|subst|pred\" \& pos!=verb)]\{,5\} [base=nie)^* [pos=ppas] | [(base=nie)^* [pos=ppas] [(pos!=\\\"adj|conj|comp|interp|subst|pred\" \& pos!=verb)]\{,5\} [base=nie)^* [(pos=praet \& base=być) [(base=by \& pos=qub) [pos=aglt]]]$

był [być:praet:sg:m1:imperf]by [by:qub] przerażony [przerazić:ppas:sg:nom:m1:perf:aff]

zobowiązany [zobowiązać:ppas:sg:nom:m1:perf:aff] był [być:praet:sg:m1:imperf]by [by:qub]

był [być:praet:sg:m3:imperf]by [by:qub] pokazywany [pokazywać:ppas:sg:nom:m3:imperf:aff]

było [być:praet:sg:n:imperf]by [by:qub] przyjęte [przyjąć:ppas:sg:nom:n:perf:aff]

były [być:praet:pl:m3:imperf]by [by:qub] selekcjonowane [selekcjonować:ppas:pl:nom:m3:imperf:aff]

nie [nie:qub] był [być:praet:sg:m1:imperf]by [by:qub]m [być:aglt:sg:pri:imperf:nwok] wpisany [wpisać:ppas:sg:nom:m1:perf:aff]

METRYCZKA

typ mowy:

rodzaj:

gatunek:

- ☐ akta sejmikowe
- ☐ bajki
- ☐ dedykacje
- ☐ dialogi
- ☐ diariusze sejmowe
- ☐ emblematy
- ☐ encyklopedie
- ☐ epitafia
- ☐ epitalamia
- ☐ fraszki i epigramaty
- ☐ herbarze
- ☐ instrukcje
- ☐ interczy
- ☐ inwentarze
- ☐ jasełka
- ☐ kalendarze
- ☐ kazania
- ☐ kolędy i pastoralki
- ☐ komedie
- ☐ konstytucje sejmowe
- ☐ kroniki
- ☐ księgi liturgiczne
- ☐ księgi sądowe
- ☐ lamenty
- ☐ modlitwy
- ☐ mowy okolicznościowe
- ☐ mowy polityczne
- ☐ ody
- ☐ opisy geograficzne
- ☐ opisy podróży
- ☐ pamiątniki
- ☐ panegiriki
- ☐ pieśni
- ☐ pisma polityczne
- ☐ podręczniki
- ☐ poematy epickie
- ☐ polemiki religijne
- ☐ poradniki
- ☐ przepisy kucharskie
- ☐ przewodniki
- ☐ przypowieści, specula (zwierciadła)
- ☐ przysłowia
- ☐ przywileje
- ☐ psalmy
- ☐ relacje
- ☐ romanse
- ☐ rozmówki do nauki języka
- ☐ satyry
- ☐ sielanki
- ☐ sonety
- ☐ 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
- ☐ zielarstwo
- ☐ zoologia
- ☐ żeglarstwo

poetyka żartu:

☐

region:

50-lecie:

uwspółcześnienie:

czasowniki... (tylko te właściwe)

13634 ZNAKÓW

[illegible]

WERSJA TESTOWA POD NKJP

niestety tylko na sucho...

python interpreter_regul_zapytan.py reguly.json przymiotnik stopien_najwyzszy_przymiotnik

```
odblokowane reguly:
  Część mowy:
    Przymiotnik, Imiesłów przymiotnikowy czynny, Imiesłów przymiotnikowy bierny
  Kategorie gramatyczne/Liczba:
    liczba pojedyncza, liczba mnoga
  Kategorie gramatyczne/Przypadek:
    mianownik, dopełniacz, celownik, biernik, narzędnik, miejscownik, wołacz
  Kategorie gramatyczne/Rodzaj:
    męski, męski osobowy, męski zwierzęcy, męski rzeczowy, żeński, nijaki
  Kategorie gramatyczne/Stopień:
    równy, wyższy, najwyższy

wybrane reguly:
  Część mowy:
    Przymiotnik
  Kategorie gramatyczne/Stopień:
    najwyższy

gotowe zapytanie:
[pos=adj && base=x] || ([pos=adja && base=x] [base="-" && base="-"]* [pos=adj &&
base=x]) || [pos="adjc||adjp" && base=x] || ([orth=najbardziej] [pos=adj]) || [
(pos=adj && degree=sup)] || ([orth=najbardziej] [pos=adj]) || ([orth=najbardziej]
[pos=adja] [base="-"]* [pos=adj]) || ([orth=najbardziej] [pos="adjc||adjp"])
['przymiotnik', 'stopien_rowny_przymiotnik', 'stopien_najwyzszy_przymiotnik']
```

WERSJA TESTOWA POD KORBĘ



- ☒ czas teraźniejszy
- ☐ czas przeszły
- ☐ czas przyszły prosty
- ☐ czas przyszły złożony

Tryb

Strona

- ☐ tryb rozkazujący
- ☐ tryb przypuszczający
- ☐ tryb rozkazujący
- ☐ tryb przypuszczający nieprzeszły
- ☐ tryb przypuszczający przeszły

1	[
2	{ "id" : "rzeczownik",
69	{ "id" : "gerundium",
131	{ "id" : "przymiotnik",
194	{ "id" : "przymiotnik niezłożony",
257	{ "id" : "przymiotnik złożony",
333	{ "id" : "czasownik",
382	{ "id" : "bezokolicznik",
441	{ "id" : "czasownik niewłasciwy",
484	{ "id" : "imiesłow przymiotnikowy czynny złożony",
551	{ "id" : "imiesłow przymiotnikowy czynny niezłożony",
618	{ "id" : "imiesłow przymiotnikowy bierny złożony",
684	{ "id" : "imiesłow przymiotnikowy bierny niezłożony",
750	{ "id" : "imiesłow przeszły",
816	{ "id" : "imiesłow przysłówkowy uprzedni",
870	{ "id" : "imiesłow przysłówkowy współczesny",
924	{ "id" : "liczebnik główny",
989	{ "id" : "liczebnik zbiorowy",
1048	{ "id" : "liczebnik przymiotnikowy",
1108	{ "id" : "liczebnik przysłówkowy",
1158	{ "id" : "zaimek osobowy",
1229	{ "id" : "zaimek zwrotny",
1300	{ "id" : "przysłówek",
1353	{ "id" : "spójnik",
1401	{ "id" : "przyimek",
1458	{ "id" : "partykula",
1506	{ "id" : "wykrzyknik",
1554	{ "id" : "skrot",
1602	{ "id" : "fragment",
1650	{ "id" : "cyfra",
1698	{ "id" : "interpunkcja",

DO ZROBIENIA:

- skuteczniejsze i bardziej zwarte kwerendy
- dodanie „metryczki”
- bardziej estetyczny interface...

mam nadzieję...

- większa modyfikowalność ze strony użytkownika; ustalanie części mowy, operatorów i tagów
- wyszukiwanie równoległe dla +dwóch dowolnych korpusów (również z uwzględnieniem różnic w językach zapytań i tagsetach)

CO DALEJ?

- części mowy vs. segmenty funkcjonalne a NLP
czasownik | strona bierna | negacja (nie) | wykładnik trybu przypuszczającego (by)
- dokładność vs. wydajność
ile mogą wytrzymać serwery?
- złoty środek
konstrukcja wieloelementowa | kilka elementów

Ja już dziękuję w pręteniyách moich za satisfácię.

*

Interpreter: Bartosz Zaborowski
GUI: Jakub Kostrzewa
Reguły: Emanuel Modrzejewski



w ramach projektu: Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do roku 1772)
(grant 0036/NPRH2/H11/81/2012)

strona korpusu (w budowie): korba.edu.pl

przykłady na podstawie:

Zasady znakowania morfosyntaktycznego w projekcie KorBa pod red. W. Gruszczyńskiego

* *Bohatyr Straszny*, Francesco Andreini, tłum. Krzysztof Piekarski, Kraków 1695, s.139 (<http://rcin.org.pl/dlibra/doccontent?id=1049>).