

Łukasz Szalkiewicz
Instytut Języka Polskiego, UW
Instytut Podstaw Informatyki, PAN



Z DOŚWIADCZEŃ RĘCZNEGO ZNAKOWANIA MILIONOWEGO PODKORPUSU NKJP LEMATYZACJA – CIEKAWY PRZYPADKI

NKJP

Publikacja

Narodowy Korpus Języka Polskiego

redakcja Adam Przepiórkowski,

Mirosław Bańko, Rafał L. Górski i

Barbara Lewandowska-Tomaszczyk,

Wydawnictwo Naukowe PWN, 2012

znakowanie

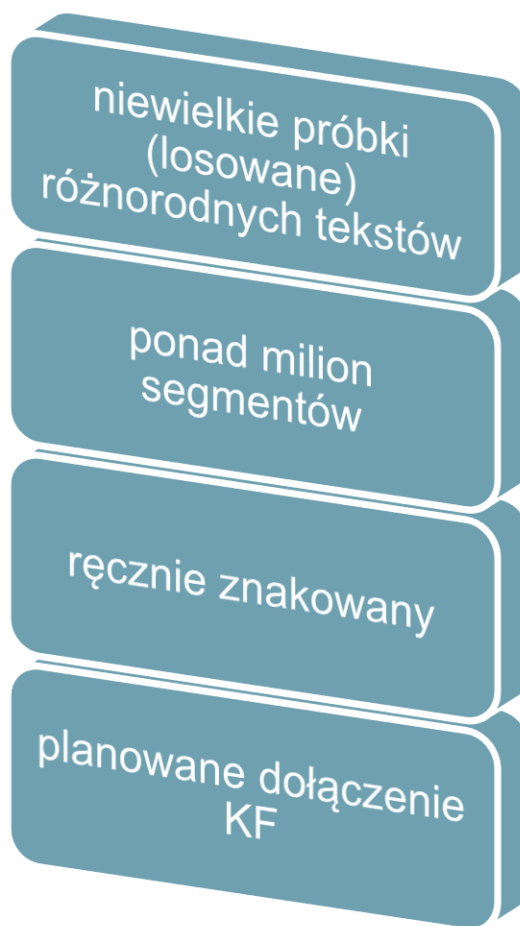
- standard - poziom morfosyntaktyczny

słowo  forma podstawowa, część mowy, kategorie morfologiczne

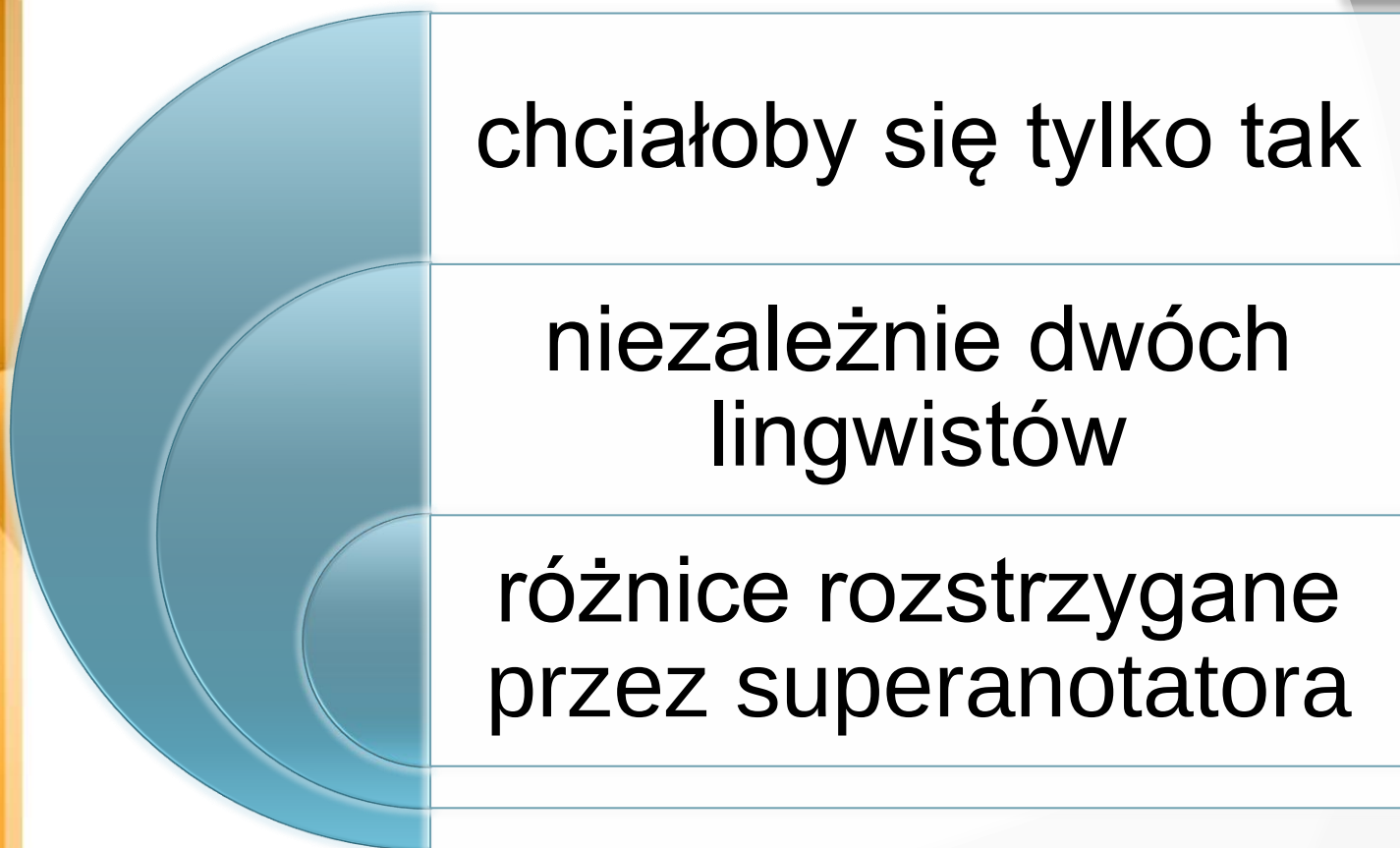
- ale też: poziom składniowy

podkorpus milonowy

<http://clip.ipipan.waw.pl/LRT>



znakowanie ręczne



znakowanie ręczne

Więc po co znakować
ręcznie?

nauka
maszyn

wartość
sama w sobie

wytrenowano
Panterę

znakowanie ręczne (Anotatornia)



tagowanie

Podstawą każdego programu tłumaczącego jest słownik.

Podstawą [podstawa:subst:sg:inst:f]

każdego [każdy:adj:sg:gen:m3:pos]

programu [program:subst:sg:gen:m3]

tłumaczącego

[tłumaczyć:pact:sg:gen:m3:imperf:aff]

jest [być:fin:sg:ter:imperf]

słownik [słownik:subst:sg:nom:m3]

pojęcia

Lemat, forma hasłowa, forma podstawowa, forma bazowa –
wybrany reprezentant

lematyzacja (hasłowanie)

fleksem

FH przynosi informację o grupowaniu
fleksemów w jeden leksem

Co było lematyzowane

Segmenty

słowa – maksymalne ciągi
znaków rozdzielone
separatorami słów („ciąg liter
pomiędzy sąsiednimi spacjami”
(Saloni, Świdziński 1998: 85))

spacja – kategoriyczny
wyznacznik segmentacji

Co było lematyzowane

2 segmenty:

*spadł|eś, długo|śmy, zjadł|by
chodź|że, znasz|li, do|ń*

3 segmenty:

*kupiła|by|m, polsko|-|francuski,
Kowalska|-|Nowakowska*

1 segment:

ping-pong, PRL-u,

Lematyzacja w NKJP

rzeczownik
(subst)

- mianownik liczby pojedynczej (mnogiej dla *plurale tantum*)
- wtórność lematyzacji
- subst/adj: *Jestem przekonany, że takie myśli mu się wtedy snuły po głowie, ale tęsknoty za **nieznanym** nie mogło wystarczyć na długo...*
- PT / nie PT: *szelki, oszczędności, media, dane, środki*

forma
deprecjatywna
(depr)

- mianownik liczby pojedynczej odpowiedniego rzeczownika

Lematyzacja w NKJP

liczebnik główny
(num)

- mianownik rodzaju męskiego rzeczowego

liczebnik zbiorowy
(numcol)

- mianownik rodzaju męskiego rzeczowego odpowiedniego liczebnika głównego

Lematyzacja w NKJP

przymiotnik
(adj)

- mianownik liczby pojedynczej rodzaju męskiego

przymiotnik
przyprzymio
tnikowy
(adja)

- mianownik liczby pojedynczej rodzaju męskiego
- *umowa polsko [polski:adja] - niemiecka*

przymiotnik
predykatyw
ny (adjc)

- mianownik liczby pojedynczej rodzaju męskiego
- *wczoraj był wesół i zdrów [zdrowy:adjc]*

przymiotnik
poprzyimko
wy (adjp)

- mianownik liczby pojedynczej rodzaju męskiego
- *agenci w mundurach i po cywilnemu [cywilny:adjp]*

Lematyzacja w NKJP

przysłówek
(adv)

- forma stopnia równego lub jedyna forma tego fleksemu
- *wydatki rosły szybciej [szybko:adv:com]*
- *pognał gdzieś [gdzieś:adv] przed siebie*

Lematyzacja w NKJP

zaimek
nietrzecioosobowy (ppron12)

- mianownik
- *kto jeszcze z [nami](#) [my:ppron12:pl:inst:m1:pri] jedzie*
- *możemy pomóc [ci](#) [ty:ppron12:sg:dat:f:sec:nakc] opiekować się*

zaimek
trzecioosobowy (ppron3)

- mianownik liczby pojedynczej
- *zawiódł [go](#) [on:ppron3:sg:acc:m1:ter:nakc:npraep] zapach*

zaimek SIEBIE
(siebie)

- SIEBIE
- *choć pociągnęłaby za [sobą](#) [siebie:siebie:inst] pewne ofiary*

Lematyzacja w NKJP

forma
nieprzeszła
(fin)

- bezokolicznik
- *już się nie zastanawiają*
[zastanawiać:fin:pl:ter:imperf]

forma
przyszła
BYĆ
(bedzie)

- BYĆ
- *jak długo tu będę* [być:bedzie:sg:pri:imperf]

aglutynant
BYĆ (aglt)

- BYĆ
- *podniosłam telefon i zostawiła m*
[być:aglt:sg:pri:imperf:nwok]

pseudonimie
słów (praet)

- bezokolicznik
- *prokurator postawił* [postawić:praet:sg:m1:perf] *mu*
zarzut

Lematyzacja w NKJP

rozkażnik
(impt)

- bezokolicznik
- odbierzmy [odebrać:impt:pl:pri:perf] telefon

bezosobnik
(imps)

- bezokolicznik
- *dopóki nie* udowodniono [udowodnić:imps:perf] winy

bezokolicznik
(inf)

- bezokolicznik
- *możemy* pomóc [pomóc:inf:perf] ci

odśłownik
(ger)

- bezokolicznik
- *podczas* patroszenia [patroszyć:ger:sg:gen:n:imperf:aff] dzika

Lematyzacja w NKJP

imiesłów
przysłówko
wy
współczesn
y (pcon)

- bezokolicznik
- *najkrócej rzecz ujmując [ujmować:pcon:imperf]*

imiesłów
przysłówko
wy uprzedni
(pant)

- bezokolicznik
- *który drzemie przysiadłszy [przysiąść:pant:perf] na dzwonie*

imiesłów
przymiotnik
owy czynny
(pact)

- bezokolicznik
- *jest zagadnieniem dotyczącym [dotyczyć:pact:sg:inst:n:imperf:aff] nas wszystkich*

imiesłów
przymiotnik
owy bierny
(ppas)

- bezokolicznik
- *ta opłata została zrównana [zrównać:ppas:sg:nom:f:perf:aff] z taryfą międzymiastową*

Lematyzacja w NKJP

winien
(winien)

- forma męska liczby pojedynczej
- powinna [powinien:winien:sg:f:imperf] *ona rozważyć*

predykatyw
(pred)

- jedyna forma tego fleksemu
- *na meblowej wystawie* można [można:pred] *oglądać dokonania*
- predykatywy a przysłówki

Lematyzacja w NKJP

przyimek (prep)

- forma niewokaliczna lub jedyna forma tego fleksemu
- *nie może we [w:prep:loc:wok] własnym zakresie*

spójnik
współrzędny
(conj)

- jedyna forma tego fleksemu
- *bez względu na indywidualne zapatrywania i [i:conj] poglądy polityczne*

spójnik
podrzędny
(comp)

- jedyna forma tego fleksemu
- *wyobrazić sobie, że [że:comp] może być inaczej*

Lematyzacja w NKJP

wykrzyknik (interj)

- jedyna forma tego fleksemu
- Ha [ha:interj], *już go widzę!*
- (O) *Jezu* – czym jest *Jezu* (interj czy subst)? (różna forma hasłowa: *Jezu*, *Jezus*)

burkinostka (conj)

- jedyna forma tego fleksemu
- *zostawił jednak drzwi otwarte na* oścież [oścież:burk]
- *wtórność lematyzacji:*
- wyciąganie dinozaurów?

kublik (qub)

- jedyna forma tego fleksemu
- *na stanowisku dowodzenia* prawie [prawie:qub] *cała jego obsługa*

Lematyzacja w NKJP

skrót (brev)

- forma hasłowa rozwinięcia
- *na przystanku przy ul
ulica:brev:pun. Rzęsnej.*

ciało obce (xxx)

- ten sam obiekt
- *Do you speak [speak:xxx] English? – zapytał.*

interpunkcja
(interp)

- ten sam obiekt
- *mnie interesują fakty, [,:interp] a nie fakty prasowe.*

niektóre nieodmienne

- ⦿ Problem: coś to jest? Ale nie: jak zlematyzować
- ⦿ ***mimo że***
- ⦿ ***Nic nie spałeś? - zatroszczyłam się.***

wyrażenia liczbowe

- ⦿ Forma hasłowa liczby zapisanej cyframi arabskimi lub liczby rzymskiej to ta sama liczba
- ⦿ *15-letni* – jeden segment z tą samą formą podstawową

skróty

- ⦿ formą hasłową skrótu jest forma hasłowa jego rozwinięcia lub całe rozwinięcie
- ⦿ *tzw.*
- ⦿ *br.*
- ⦿ inicjały
- ⦿ przewaga rozwinięcia nad formą podstawową tegoż (p., p.o., zob.)
- ⦿ oznaczenia jednostek

nazwy własne (mała a wielka litera)

- ⦿ kiedy wielka litera

*co w każdy piątek, sobotę i niedzielę spływało w oczy mieszkańców miasteczka z niewielkiego białego ekranu kina imienia **Czwartego Lutego***

nazwy własne (ortografia)

- ⦿ Nie (do końca) ustalona ortografia: *Eryk Kuna na pewno nie był dobrym żołnierzem **wehrmachtu**.*
- ⦿ *Internet*
- ⦿ *Hajdegera się naczytałeś |? Nawet w Koperfilda uwierzę - przyznał Bednarz, kiwając głową w kierunku charta.*

formy błędne

- ⦿ żeby nie. nie weszła gdzieś gdzie nie potrzeba. żeby coś nie **ściągała**
[ściągnąć:praet:sg:f:perf]
- ⦿ ręce opadają i cokolwiek jeszcze może nie ? – i majtki **spadywują**
[spadywać:fin:pl:ter:imperf]

jąkania, powtórzenia

- ⊙ *no ja moje bokserki to **prze** przelecę żelazkiem po praniu*
[przelecieć:fin:sg:pri:perf]
- ⊙ *tak bo to jest **prze** przepastne*
[przepastny:adj:sg:nom:n:pos]
- ⊙ – *zaraz poczekamy **po** poczytamy*

formy urwane (nie powtórzenia)

- ⊙ – to nie sejm się zajmuje prawdą historyczną – **prze.** prawdą historyczną się zajmują historycy
[\[przepraszać:fin:sg:pri:imperf\]](#)
- ⊙ nie mój kum mój **s** y kumpel i mój brat cioteczny zawsze bez hamulców jeżdżą.
[\[***:brev:npun\]](#)
- ⊙ tego nie słyszałam no bank przy banku bezsens przecież tych banków to tyle w tym Kaliszu **naro..** [\[narobić:praet:sg:n:perf\]](#)
- ⊙ *niektórzy się **poprzebie** pozamieniali koszulkami i teraz nie wiadomo no.. kto jest kto..* [\[poprzebierać:praet:pl:m1:perf\]](#)

Formy skrócone (nie powtórzenia)

- ◎ *Myślę, że sceny, w których "mięsozaurus" zjada człowieka lub pokazana jest odgryziona ręka czy resztki kozy spadające na szybę samochodu, nie są przeznaczone dla **pięcio-** czy ośmiolatka.*
[\[pięciolatek:subst:sg:gen:m1\]](#)
- ◎ wulgaryzmy

wykrzykniki „przeciągane”

- pojedyncze litery dla wykrzykników mieszanych, czyli „ha” dla *haaa*, „bum” dla *buuum*, „ech” dla *eeeech*, „dryń” dla *drrrrryyyń* itp.
- jedno powtórzenie dla wygłosowej spółgłoski w wykrzyknikach typu *wrrrrr*, *psssssss*, *bzzzz*, czyli: „wrr”, „pss” i „bzz”, chyba że sam wykrzyknik już jest krótszy, np. *ts*.
- Ale problematycznie:
hmmmmmm do „hm”, a nie „hmm”
syk węża *sssssss* – „ss” - jakoś dziwnie wygląda, więc może „sss”
Ledwie się wymówi teraz „trr” - ona już te swoje „*khhrrr*”, „*khhrrr*”.

wykrzykniki „przeciągane”

- pojedyncze litery dla samych samogłosek, czyli "e" dla "eeeeee", "au" dla "aaaauuuu" itp.
- *Gdyby nawet trafił przypadkiem na taki hotel, gdzie nikt nie powitałby go wylewnym „**Aaa**, moje uszanowanko!*
- *To co ja mu miałem odpowiedzieć? **Eeee!** Co? **Ueee, mmm!** Proszę jeszcze nie zamykać ust! Zaraz dam panu popłukać*

emotikonki „przeciągane”

- ⦿ *próbowano znaleźć wariant podstawowy*
- ⦿ *:-)))))) → :-)*

formy wyrazowe od 2 równoznacznych leksemów

- ⊙ *biegł* → BIEC albo BIEGNAĆ; *doprzął* → DOPRZAĆ albo DOPRZĘGNAĆ
- ⊙ *patrzą* → PATRZEĆ albo PATRZYĆ; *kaszlą* → KASZLEĆ albo KASZLAĆ; co wybrać?
- ⊙ *wiekszy* → DUŻY albo WIELKI
Teraz jesteś na to za mały, ale kiedy będziesz większy... → duży
Spośród żyjących Polaków największym jest Adam Małysz → wielki
Co w sytuacjach wątpliwych? Np.
Większe wątpliwości - wątpliwość może być, zdaje się, i duża, i wielka.

zapisy oddające wymowę – sylabizacje

- ◎ *Zadzwońę natychmiast do specjalisty od wnętrz i przyznam się, że to ja -- Melchior Faust -- urodzony w Dniu Zadusznym, w roku narastania fali — **za-mor-do-wa-łem** w wannie anatomiczną doskonałość o pretensjonalnym imieniu Julia.*
[\[zamordować praet:sg:m1:perf\]](#)
- ◎ - *Nic dziwnego! ■ Moja babunia - Parzęczewiczówna de domo - **Pa-rzędze-wi-czó-wna!** Państwu to nic nie mówi?*

zapisy oddające wymowę – inne przypadki

- ◎ *Jak to nie? ■ Pan od razu spytał **cemu** zamiast czemu.*
- ◎ *Zdrhhaajcy:*
 - [[zdrajca:subst:pl:nom:m1](#)],
- ◎ *dłuuugo:*
 - [[długo:adv:pos](#)]

gwara

- ◎ A te baby **jednygo** jeden
adj:sg:gen:m1:pos się nie **trzymio**
trzymać fin:pl:ter:imperf - zauważył
trzeci z chłopów, patrząc znacząco
na obydwu inteligentów
- ◎ Nie słuchajcie wójta i się nie **bójta!**
[bać:impt:pl:sec:imperf]

zabiegi stylistyczne

- ⦿ *I poszły, **duszając rupkami**.*
- ⦿ wprowadzamy czasownik *duszać* i rzeczownik *rupka*

Podwójne odniesienie płciowe

- ⊙ *zawiodł(a)m*
- ⊙ *niezależny(a)*
- ⊙ *pan-i*