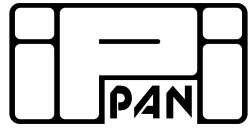
PoliMorf — otwarty słownik morfologiczny

Marcin Woliński Marcin Miłkowski Maciej Ogrodniczuk
Adam Przepiórkowski Łukasz Szałkiewicz Jan Szejko



IPI PAN, 5 grudnia 2011

Plan prezentacji

Plan prezentacji

Projekt CESAR

Central and South-East European Resources:

- projekt finansowany ze środków:
  - Komisji Europejskiej (program CIP ICT-PSP) – 50%,
  - Ministerstwa Nauki i Szkolnictwa Wyższego – 40%,
  - własnych IPI PAN – 10%.
- uruchomiony 1 lutego 2011 r. (do 31 stycznia 2013),
- wspierający inicjatywę META-NET (Multilingual Europe Technology Alliance Network of Excellence)
  `http://www.meta-net.eu`.

`http://www.cesar-project.net`

Cele projektu CESAR

Dokumenty i współpraca:

- raport o języku polskim zawierający spis dostępnych produktów, usług, technologii, zasobów; identyfikacja głównych ośrodków (badawczych, przemysłowych, rządowych, opiniotwórczych), programów, standardów i praktyk,
- współpraca z innymi projektami partnerskimi (META-NORD, METANET4U), projektem META-NET, ośrodkami krajowymi.

Zasoby i narzędzia:

- uzupełnienie zasobów i narzędzi językowych dla polszczyzny o rodzaje narzędzi dostępnych dla innych języków,
- włączenie polszczyzny w ogólnoeuropejską infrastrukturę META-SHARE (`http://www.meta-share.eu`).

Partnerzy w projekcie CESAR

Lista współpracowników:

- Hungarian Academy of Sciences, Research Institute for Linguistics – **koordynator**,
- Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics,
- University of Zagreb, Faculty of Humanities and Social Sciences,
- IPI PAN,
- Uniwersytet Łódzki,
- University of Belgrade, Faculty of Mathematics,
- Institut Mihajlo Pupin,
- Institute for Bulgarian Language, Bulgarian Academy of Sciences,
- Institute of Linguistics, Slovak Academy of Sciences.

Bieżący stan prac

Początek grudnia 2011:

W ramach pierwszej transzy projektu (następne w lipcu 2012 i styczniu 2013 r.) udostępniliśmy w repozytorium META-SHARE 8 polskich zasobów:

- korpus sejmowy,
- Słowosieć (plWordnet) w wersji 1.5,
- Nerf – narzędzie do rozpoznawania nazw własnych,
- milionowy podkorpus NKJP,
- słownik nazw własnych (gazetteer),
- korpusy audiotekstowe LUNA.PL i LUNA-WOZ.PL,
- wstępną wersję słownika morfologicznego PoliMorf.

PoliMorf = SGJP + Morfologik

Udostępnienie na licencji FreeBSD danych słowników źródłowych:

- słownika Morfologik,
- SGJP,
- wyniku scalenia danych fleksyjnych SGJP i Morfologika.

Plan prezentacji

Zygmunt Saloni
Włodzimierz Gruszczyński
Marcin Woliński
Robert Wołosz

Słownik gramatyczny języka polskiego

Wiedza Powszechna 2007

SGJP

Słownik gramatyczny języka polskiego jest projektem z długą historią:

- w latach 80-tych: analiza informacji gramatycznej w *Słowniku języka polskiego* Doroszewskiego,
- W. Gruszczyński, *Fleksja rzeczowników pospolitych we współczesnej polszczyźnie pisanej*, 1989 Wrocław,
- J. Tokarski, *Schematyczny indeks a tergo polskich form wyrazowych*, w opracowaniu Z. Saloniego, 1993 Warszawa,
- Robert Wołosz przygotował elektroniczną postać listy haseł SJPDor uzupełnionej o informację gramatyczną,
- Z. Saloni *Czasownik polski*, 2001 Warszawa.

Części składowe formy fleksyjnej

		rdzeń	zak		efobaz	char.fl.	funkcja
subst		mag	a		sg:gen	m1	D. l.p.
subst		mag	a		sg:gen	m1	B. l.p.
subst		sąg	a		sg:gen	m3	D. l.p.
adj		bia	łego		2		st. r. D. l.p.
adjcom		biels	zego		2		st. wyż. D. l.p.
adjcom	naj	biels	zego		2		st. najw. D. l.p.
v		czyta	ł	a	8	nd	...
v		czyta	ł	am	8	nd	...
v	będę␣	czyta	ł	a	8	nd	...
ppas		czyta	n	y	10	nd	M. l.p. poz.
ppas	nie	czyta	n	y	10	nd	M. l.p. neg.
		forma bazowa					

Morfeusz SGJP

- Jedna z możliwych form pochodnych danych SGJP jest używana w analizatorze morfologicznym Morfeusz SGJP.
- Przez dłuższy czas trwała dyskusja o sposobie licencjonowania, a program był dostępny tylko dla krewnych-i-znajomych.
- Od niedawna lista form używana w Morfeuszu jest dostępna na bardzo permisywnej licencji BSD.
- Morfeusz SGJP został użyty do oznakowania Narodowego Korpusu Języka Polskiego.

Dane SGJP w analizie morfologicznej

Zalety:

- Słownik prezentuje spójny metodologicznie opis fleksji o wysokim poziomie formalizacji.
- Dla większości materiału osiągnął już wysoką jakość opisu.
- Obszerny i żywy (wkrótce II wydanie).
- Możliwość dostosowania zestawu znaczników do potrzeb.

Wady:

- Niesie dziedzictwo przestarzałych leksemów z SJPDor (tzw. dynozaury).
- Wewnętrzna organizacja danych jest dość złożona.

Plan prezentacji

Morfologik

Geneza

Morfologik powstał w 2006 roku na potrzeby korektora gramatyczno-stylistycznego LanguageTool.

- Na wolnej licencji (LGPL, Creative-Commons ShareAlike, Mozilla Public License...)
- Oparty na polskim słowniku ispella,
- Morfologik powstał poprzez napisanie tabeli konwersji słownika ispella na znaczniki morfosyntaktyczne.

Morfologik

Historia danych Morfologika

- pliki tzw. dawnego słownika do ispella (Mirosław Prywata, Piotr Gackiewicz, Włodzimierz Macewicz),
- `sjp.pl`, którego pierwsza wersja (słownik alternatywny) powstała na bazie powyższego (Marek Futrega),
- program Waspell, pierwszy zawierający znaczniki (Zbigniew Płotnicki).

Morfologik

Podstawą Morfologika były tzw. flagi ispella, czyli warunkowe operacje zastępowania ciągów znaków na początku i na końcu form. Autorzy ispella, na szczęście, nazywali te operacje w sposób sensowny gramatycznie, dlatego można było wykorzystać tę regularność.

Fragment zasad konwersji

```
OSTos O ów owa ów subst:sg:gen:m
OSTos O ów owa ów subst:sg:gen:m1
OSTos O ów owem ów subst:sg:inst:m
```

Morfologik

Zalety Morfologika:

- Do niedawna jedyny wolnodostępny słownik, stosowany w wielu projektach NLP (np. Pelcra)
- Obszerny
- Aktywnie rozwijany
- Znaczniki morfosyntaktyczne w stylu korpusu IPI

Morfologik

Wady Morfologika:

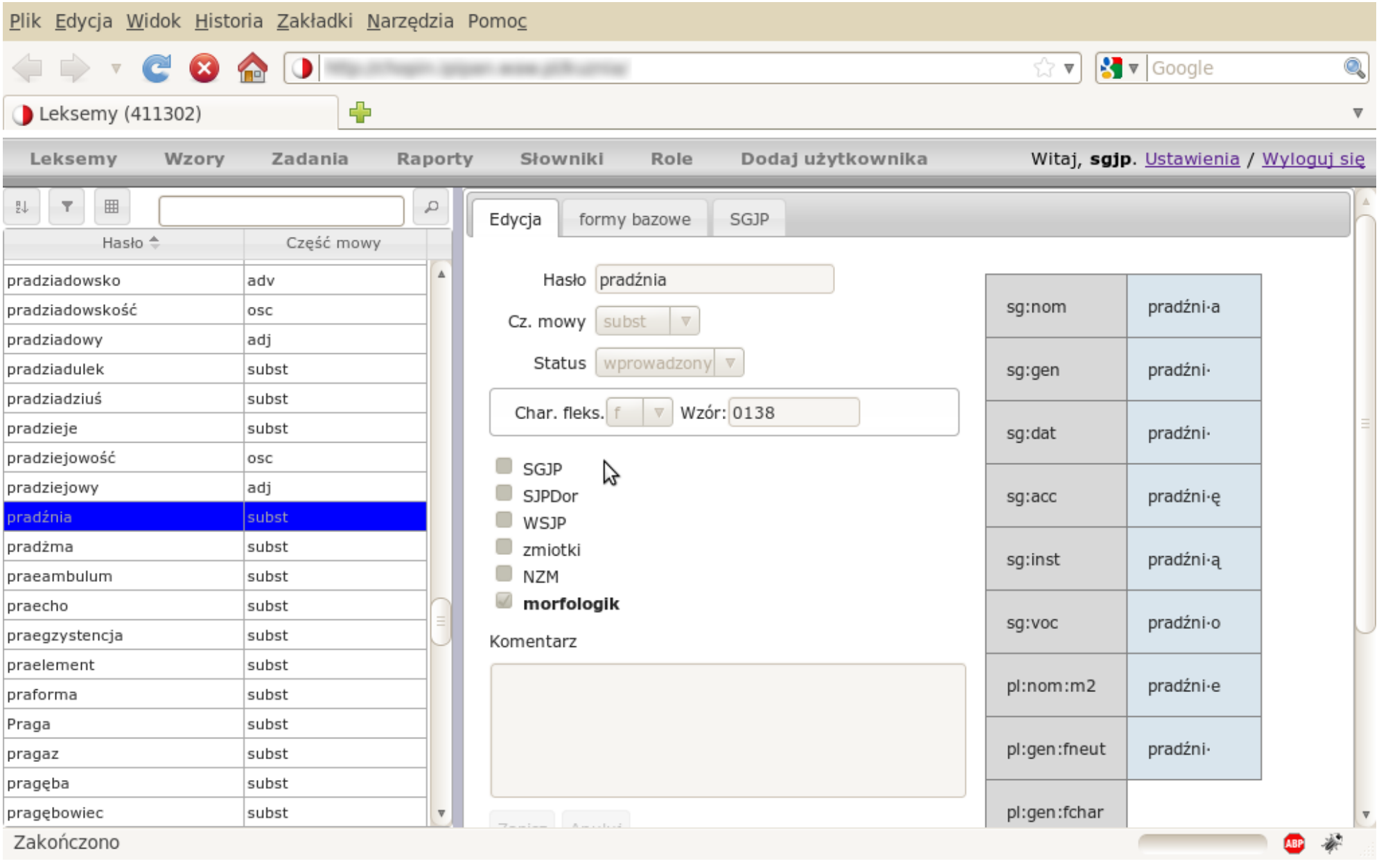
- Wyrazy, które nie były opisane zestawem tzw. flag w słowniku ispella, trzeba było dopisywać ręcznie.
- W szczególności nie ma możliwości rozróżniania rodzajów męskich ze względu na synkretyzm form.
- Marcin Miłkowski nie poprawiał słownika odpowiednio szybko.
- Duży bałagan panuje w znacznikach, nieścisły format.

Plan prezentacji

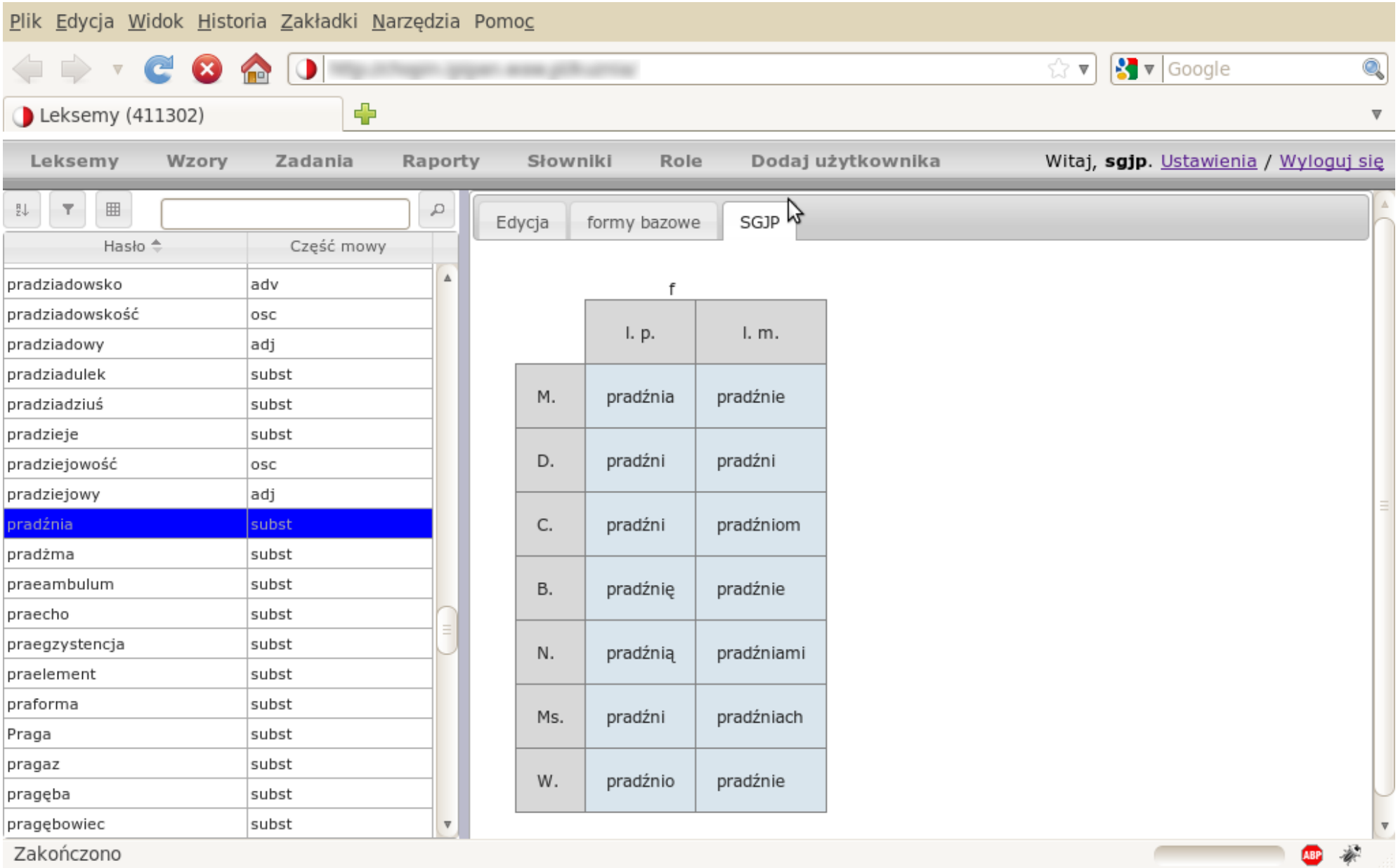
Kuźnia

- webowe środowisko pracy zespołowej nad słownikami fleksyjnymi,
- stworzone specjalnie dla projektu Cesar,
- umożliwia pracę nad wieloma słownikami, dzięki czemu
  - zachowamy tożsamość słowników składowych,
  - a także będzie można tworzyć słowniki specjalistyczne,
- jeszcze w trakcie opracowania.

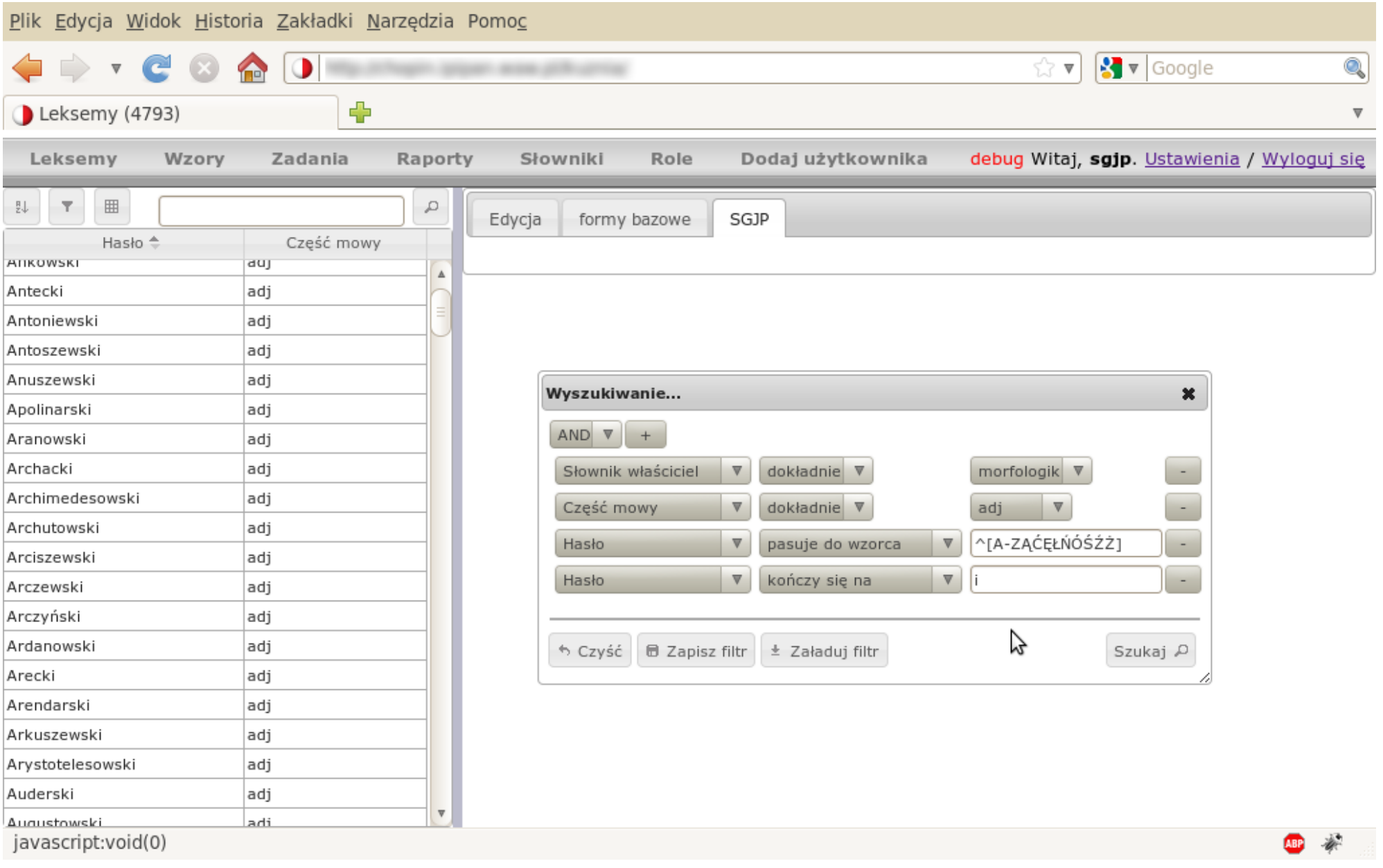
Kuźnia


Plik Edycja Widok Historia Zakładki Narzędzia Pomoc
Google
Leksemy (411302)
Leksemy Wzory Zadania Raporty Słowniki Role Dodaj użytkownika
Witaj, sgjp. Ustawienia / Wyloguj się
Hasło
Część mowy
pradziadowsko adv
pradziadowskość osc
pradziadowy adj
pradziadulek subst
pradziadziuś subst
pradzieje subst
pradziejowość osc
pradziejowy adj
pradźnia subst
pradźma subst
praeambulum subst
praecho subst
praegzystencja subst
praelement subst
praforma subst
Praga subst
pragaz subst
pragęba subst
pragębowiec subst
Edycja
formy bazowe
SGJP
Hasło pradźnia
Cz. mowy subst
Status wprowadzony
Char. fleks. f
Wzór: 0138
SGJP
SJPDor
WSJP
zmiotki
NZM
morfologik
Komentarz
sg:nom pradźni·a
sg:gen pradźni·
sg:dat pradźni·
sg:acc pradźni·ę
sg:inst pradźni·ą
sg:voc pradźni·o
pl:nom:m2 pradźni·e
pl:gen:fneut pradźni·
pl:gen:fchar
Zakończono

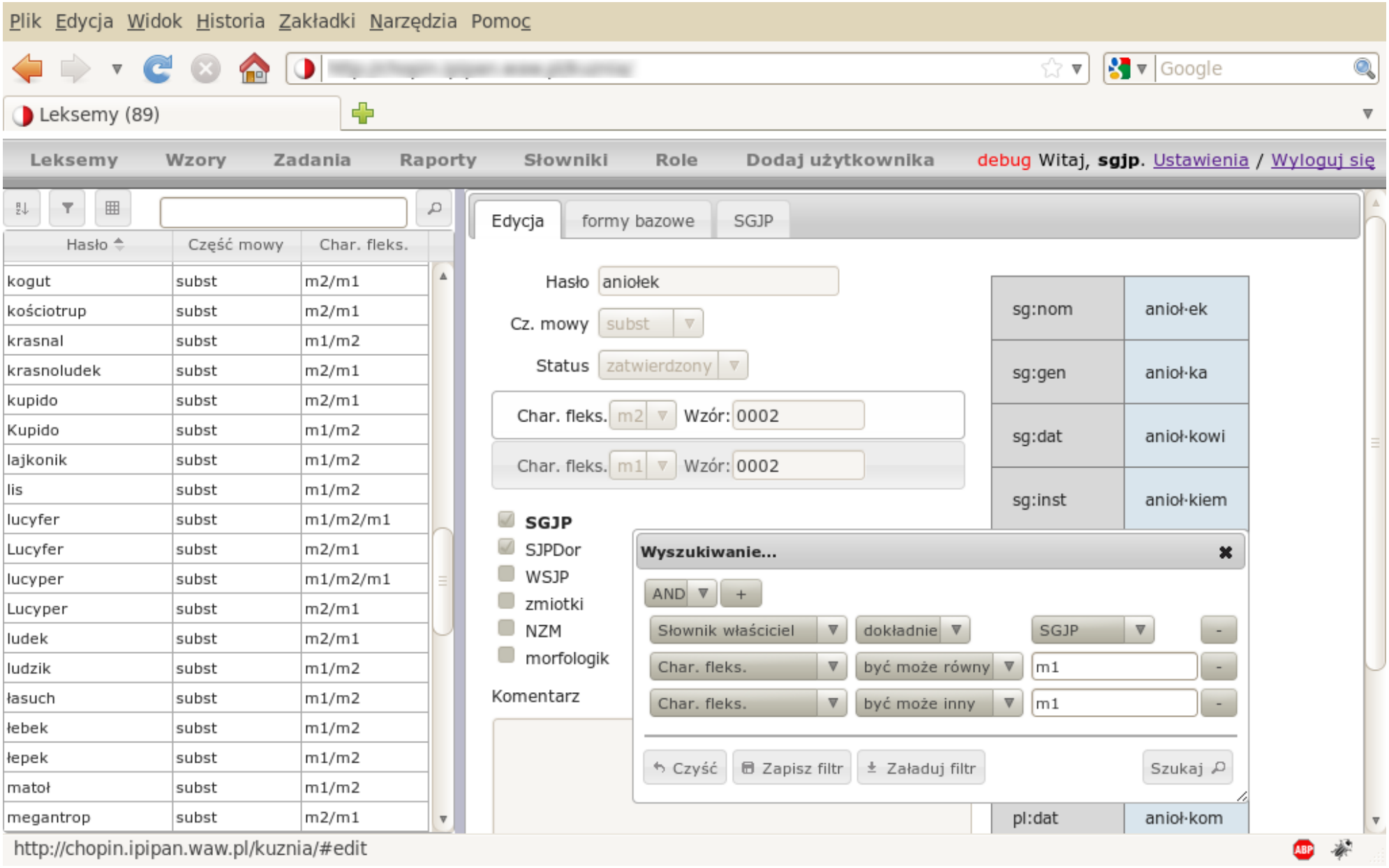
Kuźnia


Plik Edycja Widok Historia Zakładki Narzędzia Pomoc
Google
Leksemy (411302)
Leksemy Wzory Zadania Raporty Słowniki Role Dodaj użytkownika
Witaj, sgjp. Ustawienia / Wyloguj się
Hasło
Część mowy
pradziadowsko adv
pradziadowskość osc
pradziadowy adj
pradziadulek subst
pradziadziuś subst
pradzieje subst
pradziejowość osc
pradziejowy adj
pradźnia subst
pradźma subst
praeambulum subst
praecho subst
praegzystencja subst
praelement subst
praforma subst
Praga subst
pragaz subst
prageba subst
pragębowiec subst
Edycja
formy bazowe
SGJP
f
l. p.
l. m.
M. pradźnia pradźnie
D. pradźni pradźni
C. pradźni pradźniom
B. pradźnię pradźnie
N. pradźnią pradźniami
Ms. pradźni pradźniach
W. pradźnio pradźnie
Zakończono

Kuźnia


Plik Edycja Widok Historia Zakładki Narzędzia Pomoc
Google
Leksemy (4793)
Leksemy Wzory Zadania Raporty Słowniki Role Dodaj użytkownika
debug Witaj, sgjp. Ustawienia / Wyloguj się
Hasło
Część mowy
Antecki adj
Antoniewski adj
Antoszewski adj
Anuszewski adj
Apolinarski adj
Aranowski adj
Archacki adj
Archimedesowski adj
Archutowski adj
Arciszewski adj
Arczewski adj
Arczyński adj
Ardanowski adj
Arecki adj
Arendarski adj
Arkuszewski adj
Arystotelesowski adj
Auderski adj
Edycja
formy bazowe
SGJP
Wyszukiwanie...
AND
Słownik właściciel
dokładnie
morfologik
Część mowy
dokładnie
adj
Hasło
pasuje do wzorca
^[A-ZĄĆĘŁŃÓŚŹŻ]
Hasło
kończy się na
i
Czyść
Zapisz filtr
Załaduj filtr
Szukaj
javascript:void(0)

Kuźnia


Plik Edycja Widok Historia Zakładki Narzędzia Pomoc
Google
Leksemy (89)
Leksemy Wzory Zadania Raporty Słowniki Role Dodaj użytkownika
debug Witaj, sgjp. Ustawienia / Wyloguj się
Hasło Część mowy Char. fleks.
kogut subst m2/m1
kościotrup subst m2/m1
krasnal subst m1/m2
krasnoludek subst m2/m1
kupido subst m2/m1
Kupido subst m1/m2
lajkonik subst m1/m2
lis subst m1/m2
lucyfer subst m1/m2/m1
Lucyfer subst m2/m1
lucyper subst m1/m2/m1
Lucyper subst m2/m1
ludek subst m2/m1
ludzik subst m1/m2
łasuch subst m1/m2
łebek subst m1/m2
łepek subst m1/m2
matoł subst m1/m2
megantrop subst m2/m1
Edycja
formy bazowe
SGJP
Hasło aniołek
Cz. mowy subst
Status zatwierdzony
Char. fleks. m2 Wzór: 0002
Char. fleks. m1 Wzór: 0002
SGJP
SJPDor
WSJP
zmiotki
NZM
morfologik
Komentarz
sg:nom anioł·ek
sg:gen anioł·ka
sg:dat anioł·kowi
sg:inst anioł·kiem
Wyszukiwanie...
AND +
Słownik właściciel dokładnie SGJP -
Char. fleks. być może równy m1 -
Char. fleks. być może inny m1 -
Czyść Zapisz filtr Załaduj filtr Szukaj
pl:dat anioł·kom
http://chopin.ipipan.waw.pl/kuznia/#edit

Plan prezentacji

Struktura danych w SGJP i Morfologiku

SGJP

- odmiana określona parami (wzór, charakterystyka fleksyjna)
- wzór określa sposób odmiany, charakterystyka fleksyjna – układ form

Morfologik

- wiersze: wykładnik – forma podstawowa – tag, bez podziału na leksemy
- zbliżone dane są w odm.txt (dane sjp.pl): wiersze z listami wykładników leksemów, bez tagów
- oba zasoby można połączyć znakując odm.txt danymi Morfologika i dezambiguując

Schemat importu Morfologika

- znakowanie pliku odm.txt danymi z Morfologika
- dezambiguacja
- dopasowywanie wzorów i charakterystyk fleksyjnych
- utworzenie skryptu ładującego wygenerowane dane do bazy Kuźni

Dezambiguacja

- określenie, które tagi przy formie podstawowej mogą być tagami formy podstawowej
- jeśli to nie pozwala wykryć części mowy, to dezambiguacja nie udaje się
- odsiewane są tagi niepasujące do wykrytej części mowy
- dla rzeczowników wykrywany jest rodzaj i odsiewane są niepasujące tagi

Dopasowywanie wzorów – ogólny schemat

- sprawdzanie, czy takiego samego leksemu nie było już w SGJP
- filtrowanie na podstawie zakończenia formy podstawowej
- wybieranie zbioru pokrywającego zbiór wykładników form
- dla wszystkich części mowy problemami są błędy w tagach oraz nierozpoznana część form w niektórych leksemach
- w zależności od stopnia wątpliwości leksem dostaje status „kandydat” lub „wprowadzony”

Dopasowywanie wzorów – rzeczowniki

- filtrowanie wzorów na podstawie charakterystyki fleksyjnej
- wykrywanie rzeczowników jednoliczbowych: najpierw według tagów, potem według wzorów zawierających oczekiwane formy
- dopasowywanie wzorów do rzeczowników bez liczby mnogiej przez szukanie podobnych w SGJP i kopiowanie odmiany
- charakterystyka fleksyjna (rodzaj) często niemożliwa do ustalenia

Dopasowywanie wzorów – czasowniki

- w SGJP wzory dla czasowników generują zestaw form bazowych, z których wszystkie pozostałe są tworzone sufiksami
- dane z Morfologika są analizowane tak, by ustalić zbiór form bazowych, do którego następnie dopasowywane są wzory
- tagi określające charakterystykę fleksyjną (aspekt) nie są konsekwentne – patrzymy na istnienie imiesłowów przysłówkowych (z wyjątkiem czasowników bezpodmiotowych)

Dopasowywanie wzorów – nieodmienne

- oddzielna procedura
- tylko dwa możliwe wzory: jedna forma lub dwie (*nad*/*nade*)
- w Morfologiku brak rozróżnienia adv/advndm i conj/comp

Plan prezentacji

PoliMorf 0.5

pradziejowy	pradziejowy	adj:sg:nom.voc:m1.m2.m3:pos
pradziejowe	pradziejowy	adj:sg:nom.voc:n1.n2:pos
pradźnie	pradźnia	subst:pl:acc:f
pradźniom	pradźnia	subst:pl:dat:f
pradźni	pradźnia	subst:pl:gen:f
pradźniami	pradźnia	subst:pl:inst:f
pradźniach	pradźnia	subst:pl:loc:f
pradźnie	pradźnia	subst:pl:nom.voc:f
pradźnię	pradźnia	subst:sg:acc:f
pradźni	pradźnia	subst:sg:dat:f
pradźni	pradźnia	subst:sg:gen:f
pradźnią	pradźnia	subst:sg:inst:f
pradźni	pradźnia	subst:sg:loc:f
pradźnia	pradźnia	subst:sg:nom:f
pradźnio	pradźnia	subst:sg:voc:f
pradżmy	pradżma	subst:pl:acc:f

PoliMorf 0.5

	wspólne	tylko SGJP	tylko Morfologik	razem
rzeczowniki	72378	94597	51707	218682
„prawdziwe”	52723	25619	47107	125449
gerundia	16782	12938	4600	34320
-ość	2423	27282		29705
nie-. . . -ość	450	28758		29208
przymiotniki	70537	26386	32064	128987
st. równy	24433	7190	16415	48038
nie-	23737	6033	10172	39942
st. wyższy	863	178	172	1213
im. czynny	7960	6036	1763	15759
im. bierny	13544	6949	3542	24035
czasowniki	16784	12890	4666	34340
nieodmienne	9017	16694	2417	28128
razem	**168716**	**150567**	**90854**	**410137**

PoliMorf 0.5

	wspólne	tylko SGJP	tylko Morfologik	razem
nieodmienne	9017	16694	2417	28128
adv	4067	7489	2088	13644
adv *nie-*	3799	7580		11379
advcom	869	357	134	1360
advndm	122	388		510
prep	47	68	18	133
comp	23	33		56
conj	30	34	19	83
interj	9	420		429
qub	51	191	156	398
burk		134	2	136

Plan prezentacji

Dalszy rozwój PoliMorfa

- Zaimportowane dane wymagają weryfikacji i korekty.
- Zostaną wzbogacone o klasyfikację nazw własnych i kwalifikatory.
- Problemy badawczo-dyskusyjne:
  - jak opisywać skróty (w Morfeuszu SGJP opis raczej skąpy)?
  - jak opisywać jednostki typu *czterokonny*, *20-letni*, *obiadeś*, *doń*, *antystół*?

Sposoby używania PoliMorfa

- Stworzymy mechanizm pobierania z Kuźni list form dostosowanych do konkretnych zastosowań (przede wszystkim przez wybór odpowiedniego podzbioru słowników).
- Planujemy regularne wydania w miarę poprawiania i uzupełniania danych.

Sposoby używania PoliMorfa

Wątpliwości:

- Czy użytkownicy oczekują dostępności danych, czy narzędzia ich używającego? (obu)
- Czy należy generować jakieś inne formaty listy form oprócz trzykolumnowego?
- Jakie tagsety powinniśmy uwzględnić? (Morfeusz i Morfologik)

A przy okazji:

Ulepszymy Morfeusza!

- informacja o imionach, nazwiskach, geogr., innych własnych,
- oznaczać formy dawne i przestarzałe,
- zrewidować reguły łączenia subsegmentów,
- kroczek w stronę derywacji: prefiksacja?
- opcjonalnie bez segmentowania czasowników wewnątrzsłowowo,
- analiza z uwzględnieniem kasztowości,
- możliwość użycia w programach wielowątkowych,
- odmieniacz działający w sposób spójny z analizatorem (to wymaga wprowadzenia oznaczeń homonimów).

PoliMorf już dostępny

http://zil.ipipan.waw.pl/PoliMorf