Narzędzia znakowania fleksyjnego tekstów polskich: Morfeusz 2 i Concraft 2

Marcin Woliński

Zespół Inżynierii Lingwistycznej
Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Toruń, 16 listopada 2018

Wprowadzenie

Głównym tematem warsztatu będzie system znakowania fleksyjnego używany

- w analizatorze Morfeusz 2 SGJP,
- (ze zmianami) w Narodowym Korpusie Języka Polskiego: `http://nkjp.pl/poliqarp`
- (z innymi zmianami) w korpusach historycznych:
  - tekstów z XVII i XVIII w. (do 1772 r.): `http://korba.edu.pl/`
  - tekstów z lat 1830–1918: `http://korpus19.nlp.ipipan.waw.pl/`
- w narzędziu Korpusomat: `http://korpusomat.pl/`

Analizator i generator fleksyjny Morfeusz

Morfeusz, wersja 2
opracowana w Zespole Inżynierii Lingwistycznej IPI PAN
w ramach CLARIN-PL:

`http://sgjp.pl/morfeusz`

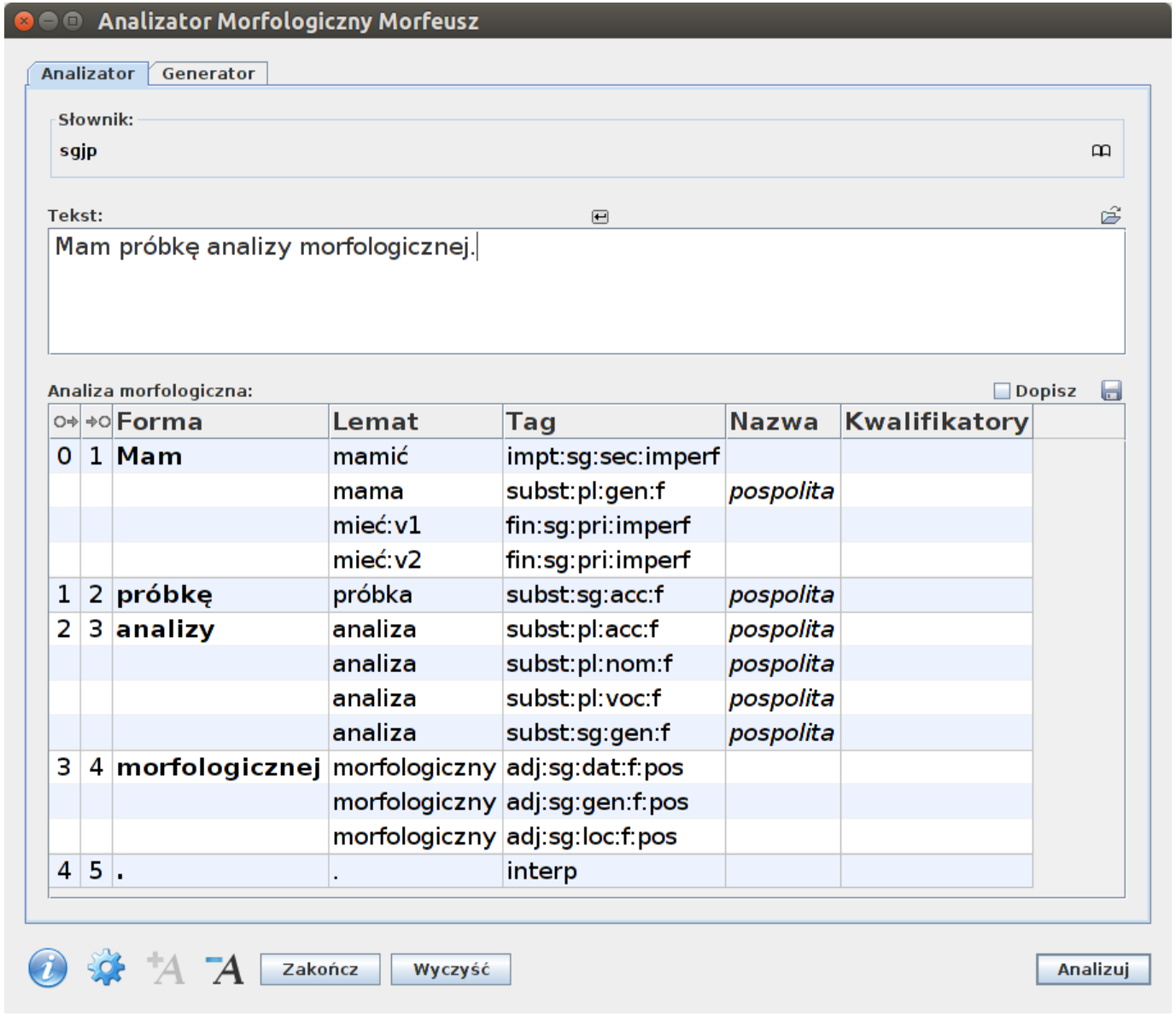
Wersja demonstracyjna on-line:

`http://sgjp.pl/morfeusz/demo`

Dystrybucja programu

- Zasadniczą postać programu stanowi moduł programistyczny, który można wbudować w tworzone przez siebie programy.
- Dla mniej technicznie ukierunkowanych użytkowników przygotowano interfejs okienkowy.
- Udostępniamy kod źródłowy i wersje skompilowane dla Linuksa, Mac OS X i Windows; 32- i 64-bitowe.
- Dodatkowe moduły umożliwiają użycie Morfeusza z poziomu Pythona, Javy i SWI-Prologu.

Interfejs okienkowy Morfeusza



		Forma	Lemat	Tag	Nazwa	Kwalifikatory
0	1	**Mam**	mamić	impt:sg:sec:imperf		
			mama	subst:pl:gen:f	*pospolita*	
			mieć:v1	fin:sg:pri:imperf		
			mieć:v2	fin:sg:pri:imperf		
1	2	**próbkę**	próbka	subst:sg:acc:f	*pospolita*	
2	3	**analizy**	analiza	subst:pl:acc:f	*pospolita*	
			analiza	subst:pl:nom:f	*pospolita*	
			analiza	subst:pl:voc:f	*pospolita*	
			analiza	subst:sg:gen:f	*pospolita*	
3	4	**morfologicznej**	morfologiczny	adj:sg:dat:f:pos		
			morfologiczny	adj:sg:gen:f:pos		
			morfologiczny	adj:sg:loc:f:pos		
4	5	**.**	.	interp		

Podstawowe pojęcia

leksem (wyraz słownikowy) abstrakcyjna jednostka języka, zbiór form wyrazowych

forma (fleksyjna) segment zinterpretowany poprzez przypisanie do leksemu i określenie jego funkcji gramatycznej

segment (wykładnik formy) jej reprezentacja w tekście

lemat umowny identyfikator leksemu, tradycyjnie równokształtny z wykładnikiem pewnej jego formy

Podstawowe pojęcia

leksem (wyraz słownikowy) abstrakcyjna jednostka języka, zbiór form wyrazowych

forma (fleksyjna) segment zinterpretowany poprzez przypisanie do leksemu i określenie jego funkcji gramatycznej

segment (wykładnik formy) jej reprezentacja w tekście

lemat umowny identyfikator leksemu, tradycyjnie równokształtny z wykładnikiem pewnej jego formy

Technicznie:

forma trójka $\langle$*segment, lemat, znacznik fleksyjny (tag)*$\rangle$

leksem zbiór form o tym samym lemacie

Podstawowe pojęcia

Analiza fleksyjna (morfologiczna) to identyfikacja wszystkich form wyrazowych, których dany segment może być wykładnikiem.

Ujednoznacznianie fleksyjne to określenie na podstawie kontekstu, jako którą z możliwych form interpretować dane wystąpienie segmentu.

Tagowanie = analiza + ujednoznacznienie

Przykład

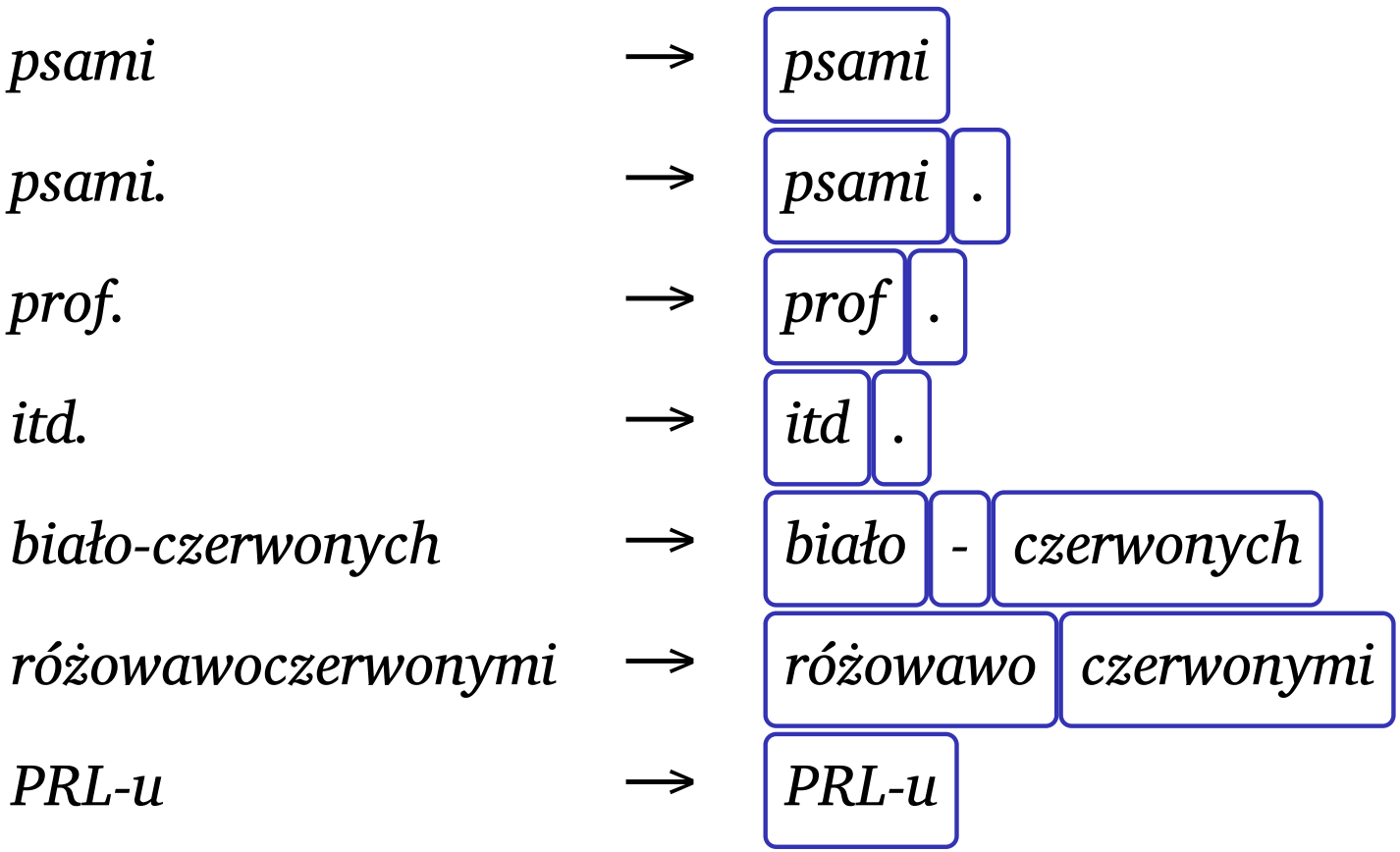
Mam	MAMA MAMIĆ MIEĆ	subst:pl:gen:f impt:sg:sec:imperf fin:sg:pri:imperf
próbkę	PRÓBKA	subst:sg:acc:f
analizy	ANALIZA	subst:sg:gen:f subst:pl:nom.acc.voc:f
morfologicznej	MORFOLOGICZNY	adj:sg:gen.dat.loc:f:pos
.	.	interp

Segmentacja

słowo maksymalny ciąg znaków nie zawierający odstępu

segment minimalny odcinek tekstu podlegający interpretacji fleksyjnej

Segmentacja



Segmentacja polszczyzny jest uwikłana słownikowo.

Segmenty krótsze od słów

- *Powiedziała, że to* ***czytaliście****.*
- *Powiedziała, że****ście*** *to* ***czytali****.*
- **Powiedziała, żeby to* ***czytaliście****.*
- *Powiedziała, żeby****ście*** *to* ***czytali****.*

Segmenty krótsze od słów

- *Powiedziała, że to* ***czytaliście****.*
- *Powiedziała, że****ście*** *to* ***czytali****.*
- **Powiedziała, żeby to* ***czytaliście****.*
- *Powiedziała, żeby****ście*** *to* ***czytali****.*
- *Świnie****ście****!*

Segmentacja form czasu przeszłego

Wariant fundamentalistyczny:

widział	WIDZIEĆ	praet:sg:m1.m2.m3:imperf
em	BYĆ	aglt:sg:pri:imperf:wok

Wariant pragmatyczny:

widziałem	WIDZIEĆ	praet:sg:m1.m2.m3:pri:imperf

Segmentacja form trybu warunkowego

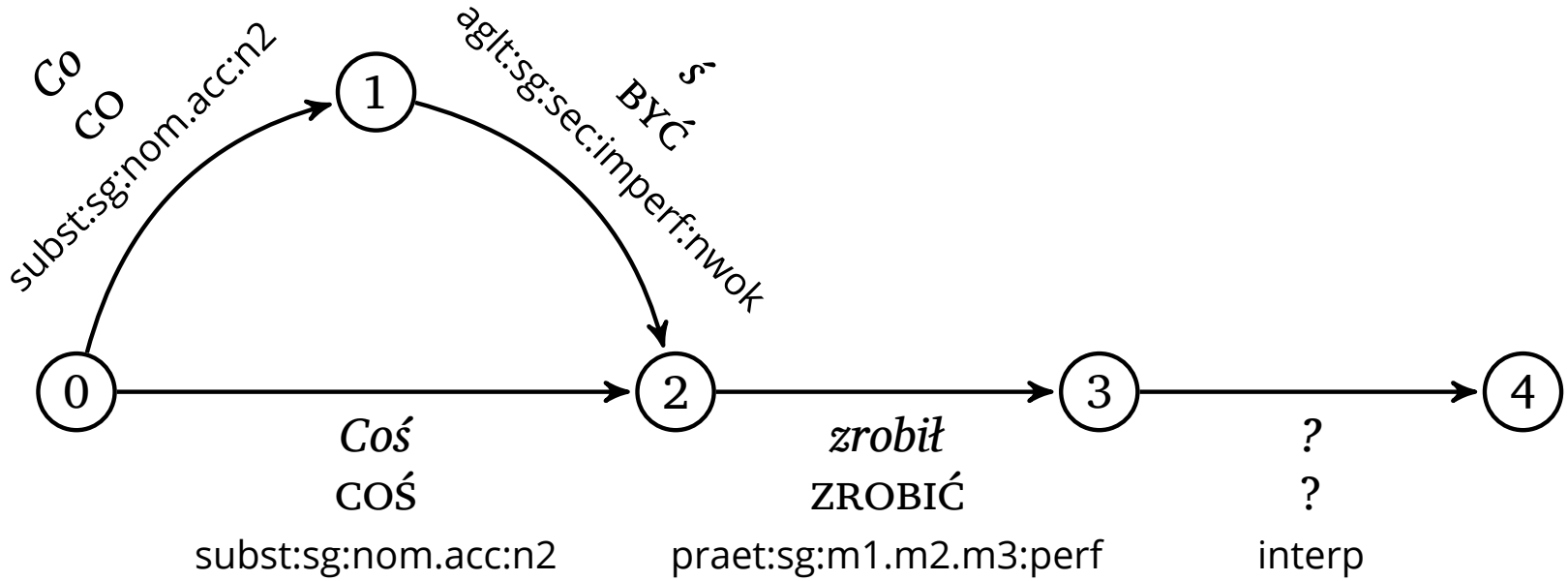
Wariant fundamentalistyczny:

widział	WIDZIEĆ	praet:sg:m1.m2.m3:imperf
by	BY	qub
m	BYĆ	aglt:sg:pri:imperf:wok

Wariant pragmatyczny (nowy znacznik cond):

widziałbym	WIDZIEĆ	cond:sg:m1.m2.m3:pri:imperf

Reprezentacja wyników analizy



Kategoria rodzaju gramatycznego

Klasy rodzajowe według Mańczaka (1956):

	m1	m2	m3	n	f
acc. sing.	*tego*		*ten*	*to*	*tę*
acc. pl.	*tych*	*te*			

Ten zbiór wartości jest używany w znakowaniu NKJP.

Kategoria rodzaju gramatycznego

Klasy rodzajowe według Saloniego (1976):

	m1	m2	m3	n1	n2	f	p1	p2	p3
acc. sing.	*tego*		*ten*	*to*		*tę*	⊥		
acc. pl.	*tych*	*te*					*tych*	*te*	
acc. pl.	*pięciu*	*pięć*		*pięcioro*	*pięć*		*pięcioro*		⊥

Ten zbiór wartości jest używany w SGJP i wcześniejszych wersjach analizatora Morfeusz.

Kategoria rodzaju gramatycznego

W bieżącej wersji Morfeusza:

- kategoria rodzaju o wartościach m1, m2, m3, f, n;
- kategoria przyrodzaju o wartościach col (zbiorowy), ncol (główny), pt (zbiorowy *plurale tantum*) stosowana wyłącznie w znacznikach rzeczowników i liczebników.

Regularność odmiany

TRYB OZNAJMUJĄCY

Czas teraźniejszy^{ndk} / *przyszły*^{dk}

lp

1.os.	**kra**dnę
2.os.	**kra**dniesz
3.os.	**kra**dnie

lm

1.os.	**kra**dniemy
2.os.	**kra**dniecie
3.os.	**kra**dną

Czas przeszły

lp

m	**kra**dł(e)	m	1.os.
ż	**kra**dła	ś	2.os.
n	**kra**dło	∅	3.os.

bezosobnik: **kra**dziono

lm

mo	**kra**dli	śmy	1.os.
nmo	**kra**dły	ście	2.os.
		∅	3.os.

TRYB ROZKAZUJĄCY

lp

2.os.	**kra**dnij

lm

1.os.	**kra**dnijmy
2.os.	**kra**dnijcie

Bezokolicznik: **kra**ść

Regularność odmiany

TRYB OZNAJMUJĄCY

Czas teraźniejszy^{ndk} / *przyszły*^{dk}

lp		lm	
1.os.	**kra**dnę	1.os.	**kra**dniemy
2.os.	**kra**dniesz	2.os.	**kra**dniecie
3.os.	**kra**dnie	3.os.	**kra**dną

Czas przeszły

lp				lm			
m	**kra**dł(e)	m	1.os.	mo	**kra**dli	śmy	1.os.
ż	**kra**dła	ś	2.os.	nmo	**kra**dły	ście	2.os.
n	**kra**dło	∅	3.os.			∅	3.os.

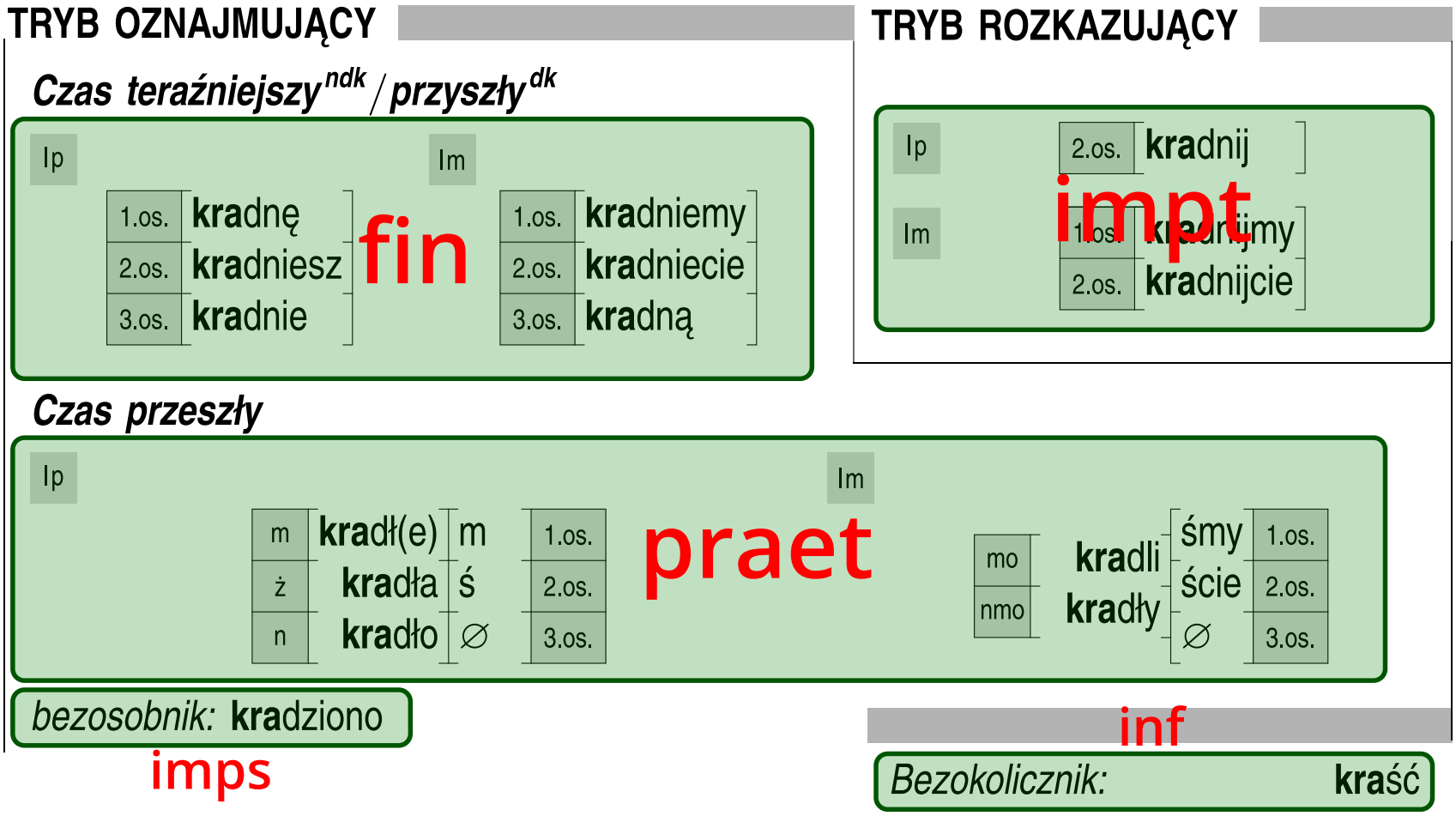
bezosobnik: **kra**dziono

TRYB ROZKAZUJĄCY

lp			lm		
2.os.	**kra**dnij		1.os.	**kra**dnijmy	
			2.os.	**kra**dnijcie	

Bezokolicznik: **kra**ść

Regularność odmiany



Fleksem podzbiór leksemu (w miarę) jednorodny ze względu na kategorie gramatyczne przysługujące formom

Tagset

Mam	MAMA MAMIĆ MIEĆ	subst:pl:gen:f impt:sg:sec:imperf fin:sg:pri:imperf
próbkę	PRÓBKA	subst:sg:acc:f
analizy	ANALIZA	subst:sg:gen:f subst:pl:nom.acc.voc:f
morfologicznej	MORFOLOGICZNY	adj:sg:gen.dat.loc:f:pos
.	.	interp

Lematyzacja

Leksem PARA:

- *Uczestnicy tańczą **parami**.*
- *Zatrucie **parami** rtęci jest praktycznie niemożliwe bez jednoczesnego poparzenia.*

Lematyzacja

Leksem PARA:

- *Uczestnicy tańczą **parami**.*
- *Zatrucie **parami** rtęci jest praktycznie niemożliwe bez jednoczesnego poparzenia.*

Leksemy ZAMEK:S1 i ZAMEK:S2:

- *Jakoś odruchowo przekręciła gałkę **zamka**, a potem nacisnęła klamkę.*
- *Na dziedzińcu **zamku** lubelskiego natrafiono na fragmenty konstrukcji zrębowej drewnianej chaty.*

Lematyzacja

- Lematy ok. 10 000 leksemów w SGJP wymagają elementu ujednoznaczniającego.
- Po dwukropku dodano oznaczenie części mowy.
  Np. leksemy PIEC:S i PIEC:V.
- Jeżeli to nie wystarczyło, dodano oznaczenie cyfrowe, np. ZAMEK:S1 (*zamka*) i ZAMEK:S2 (*zamku*); SŁAĆ:V1 (*ślę*) i SŁAĆ:V2 (*ścielę*).
- Analizator zwraca takie lematy.
- Generator dla argumentu `"piec:s"` zwróci formy odmiany rzeczownika PIEC:S, a dla argumentu `"piec"` — formy zarówno rzeczownika jak i czasownika.

Opcje programu: dopuszczanie aglutynantów

- Słownik Morfeusza niesie ze sobą również definicje sposobów łączenia segmentów.
- Dla domyślnych słowników dostępne są opcje aglutynacji „strict” i „permissive”, dopuszczające odpowiednio dołączanie cząstek aglutynacyjnych ograniczone i bardziej swobodne.
- Można zdefiniować kolejne warianty i rozpoznawać słowa typu:

  ***Potrzebowałżebyś**, pytam na koniec, tego strachu wstrętnego i bezsilnej wściekłości.* (Lem, *Przyjaciel Automateusza*)

Opcje programu: czas przeszły

Dostępne są dwa sposoby interpretowania form czasu przeszłego i trybu warunkowego:

- „split” (domyślna) — traktowane jako złożone z formy czasu przeszłego i aglutynantu,
- „composite” — traktowane jako pojedyncze segmenty (z wyjątkiem form jawnie analitycznych).

Opcje programu: wielkie litery

Morfeusz ma trzy tryby wrażliwości na wielkie litery:

- niewrażliwy — wielkie i małe litery nie wpływają na rozpoznawanie form,
- wrażliwy — *Polski* analizowane jako POLSKI i POLSKA; *polski* analizowane tylko jako POLSKI; *andrzej* — `ign`,
- wrażliwy warunkowo — jak poprzedni, ale *andrzej* zostanie zanalizowane jako forma leksemu ANDRZEJ, bo jest to jedyna interpretacja.

Informacje dodatkowe

Morfeusz 2 dodaje w wynikach analizy dwa elementy, które nie są ściśle fleksyjne:

- prostą klasyfikację nazw własnych,
- kwalifikatory.

```
Toruniu,Toruń,subst:sg:loc:m3,nazwa geograficzna,_
Marcina,Marcin,subst:sg:gen.acc:m1,imię,_
tą,ten,adj:sg:inst:f:pos,_,_
tą,ten,adj:sg:acc:f:pos,_,pot.
```

Dostępne słowniki

- Morfeusz jest dystrybuowany z dwoma słownikami:
  - SGJP (`http://sgjp.pl/`)
    - ponad 300 tysięcy leksemów
    - ponad 4 miliony wykładników form
  - Polimorf (`http://zil.ipipan.waw.pl/PoliMorf`).
- Kolejne wydania Morfeusza są generowane automatycznie przez system *Kuźnia* zarządzający pracą nad oboma słownikami.

Kompilowanie słownika

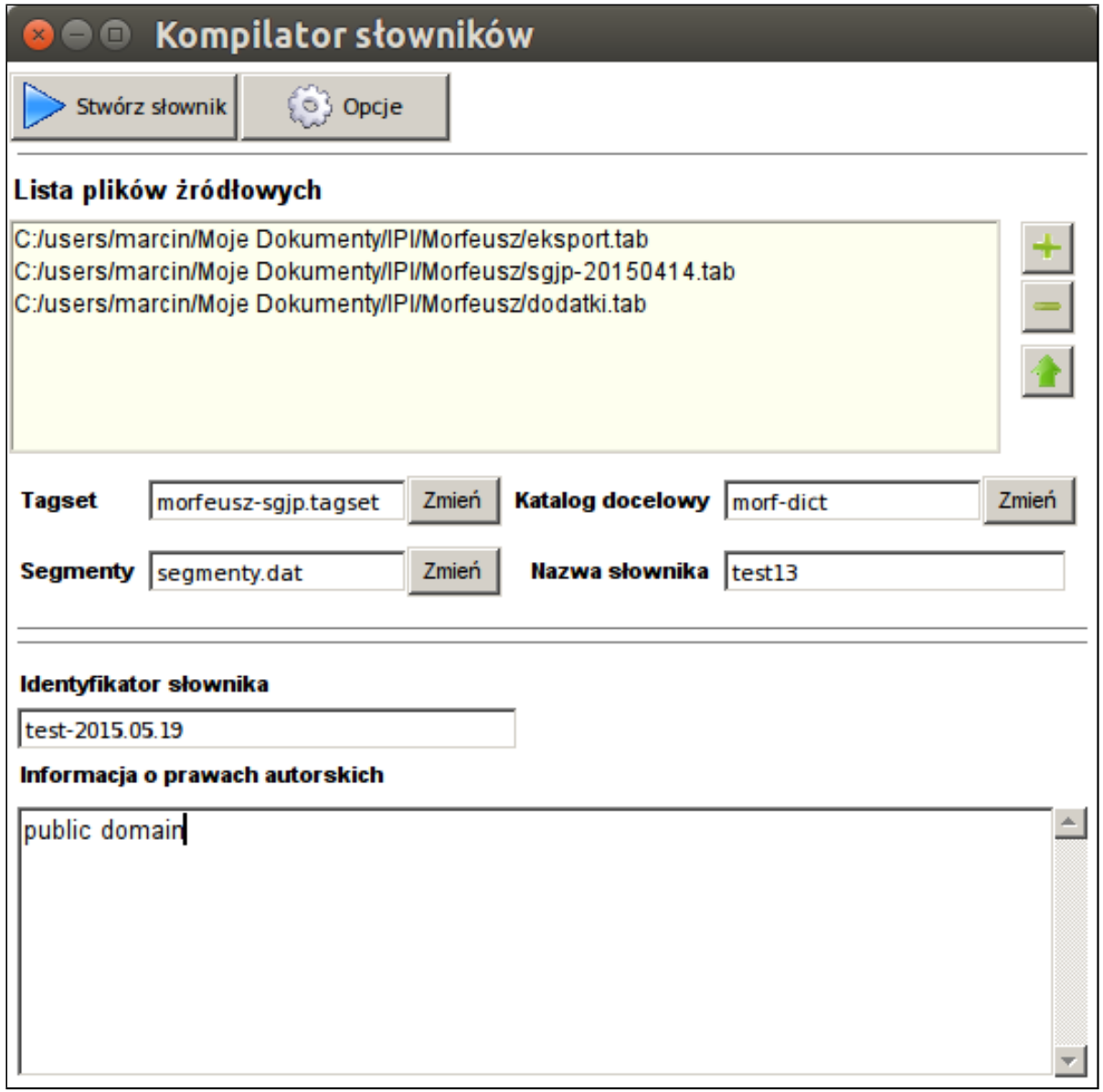
Dane wbudowywane w binarny plik słownikowy Morfeusza:

- słownik lub słowniki źródłowe,
- reguły łączenia segmentów,
- definicja tagsetu.

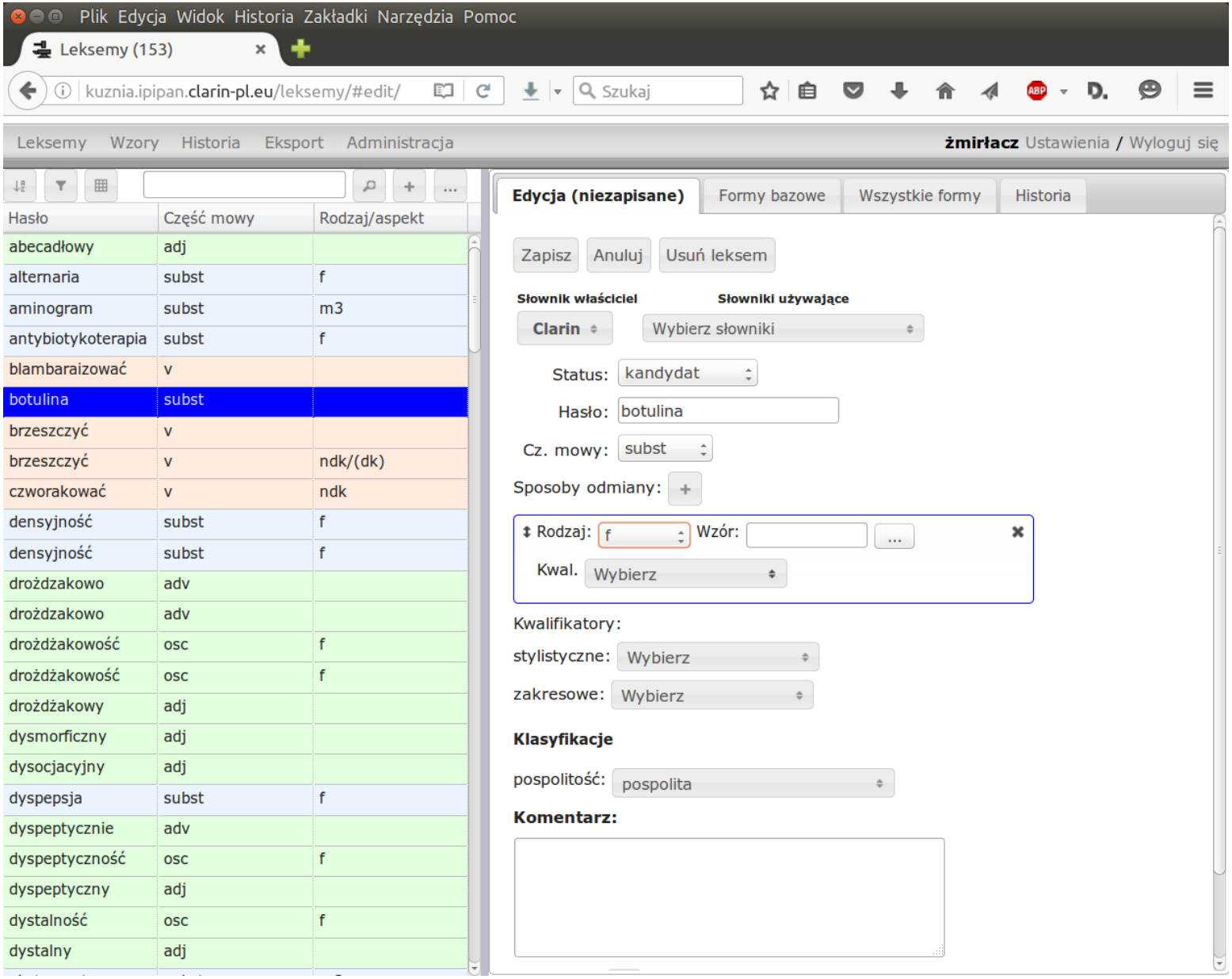
Postać źródłowa słownika

```
Gdańsk    Gdańsk  subst:sg:acc:m3  geograficzna
Gdańsk    Gdańsk  subst:sg:nom:m3  geograficzna
Gdańska   Gdańsk  subst:sg:gen:m3  geograficzna
Gdański   Gdańsk  subst:pl:nom:m3  geograficzna
Gdańskiem Gdańsk  subst:sg:inst:m3 geograficzna
funkcja   funkcja subst:sg:nom:f   pospolita
funkcjach funkcja subst:pl:loc:f   pospolita
funkcjami funkcja subst:pl:inst:f  pospolita
funkcje   funkcja subst:pl:acc:f   pospolita
funkcje   funkcja subst:pl:nom:f   pospolita
funkcje   funkcja subst:pl:voc:f   pospolita     rzad.
funkcji   funkcja subst:pl:gen:f   pospolita
funkcji   funkcja subst:sg:gen:f   pospolita
funkcjo   funkcja subst:sg:voc:f   pospolita     rzad.
funkcjom  funkcja subst:pl:dat:f   pospolita
funkcyj   funkcja subst:pl:gen:f   pospolita     arch.
```

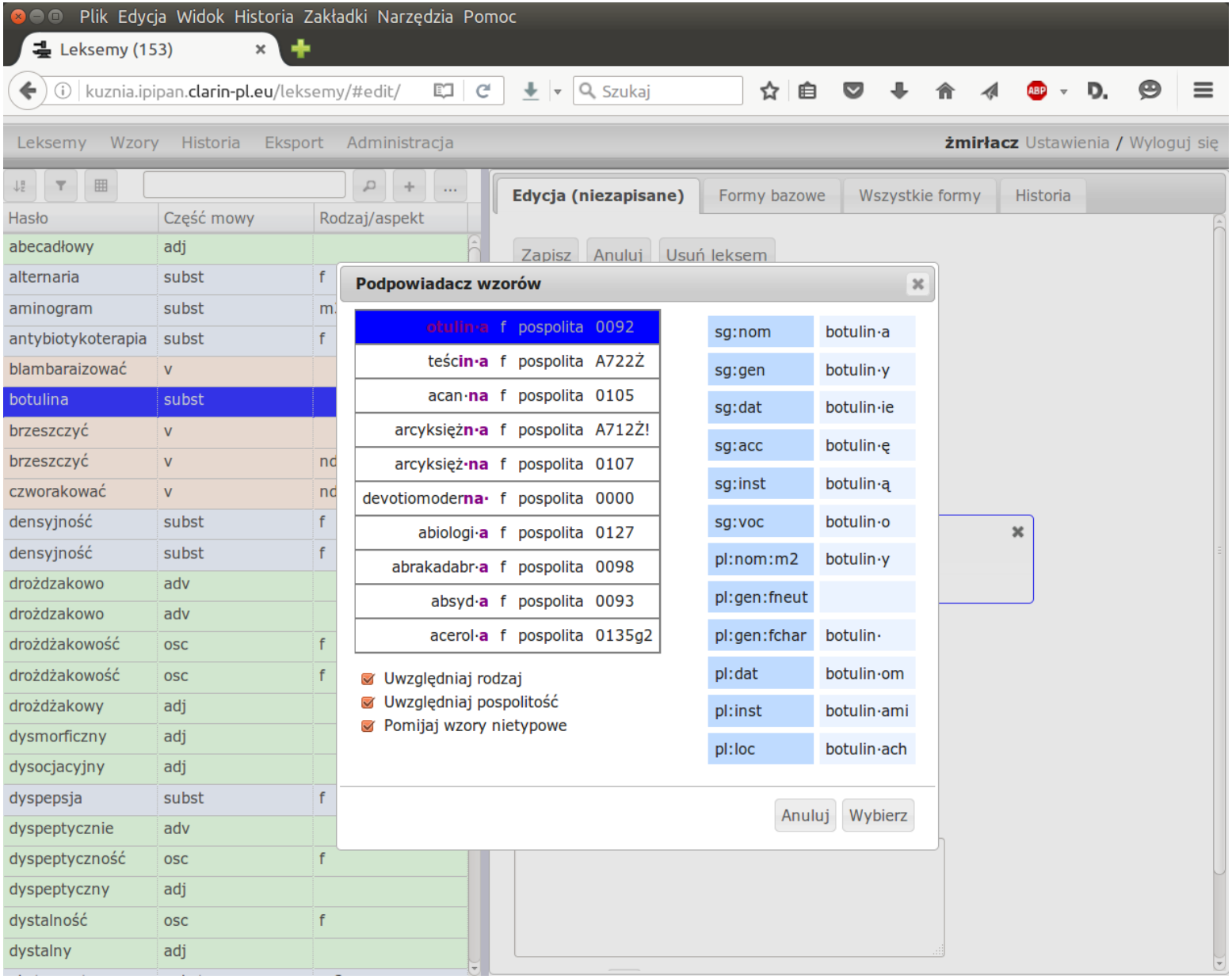
Kompilowanie słownika



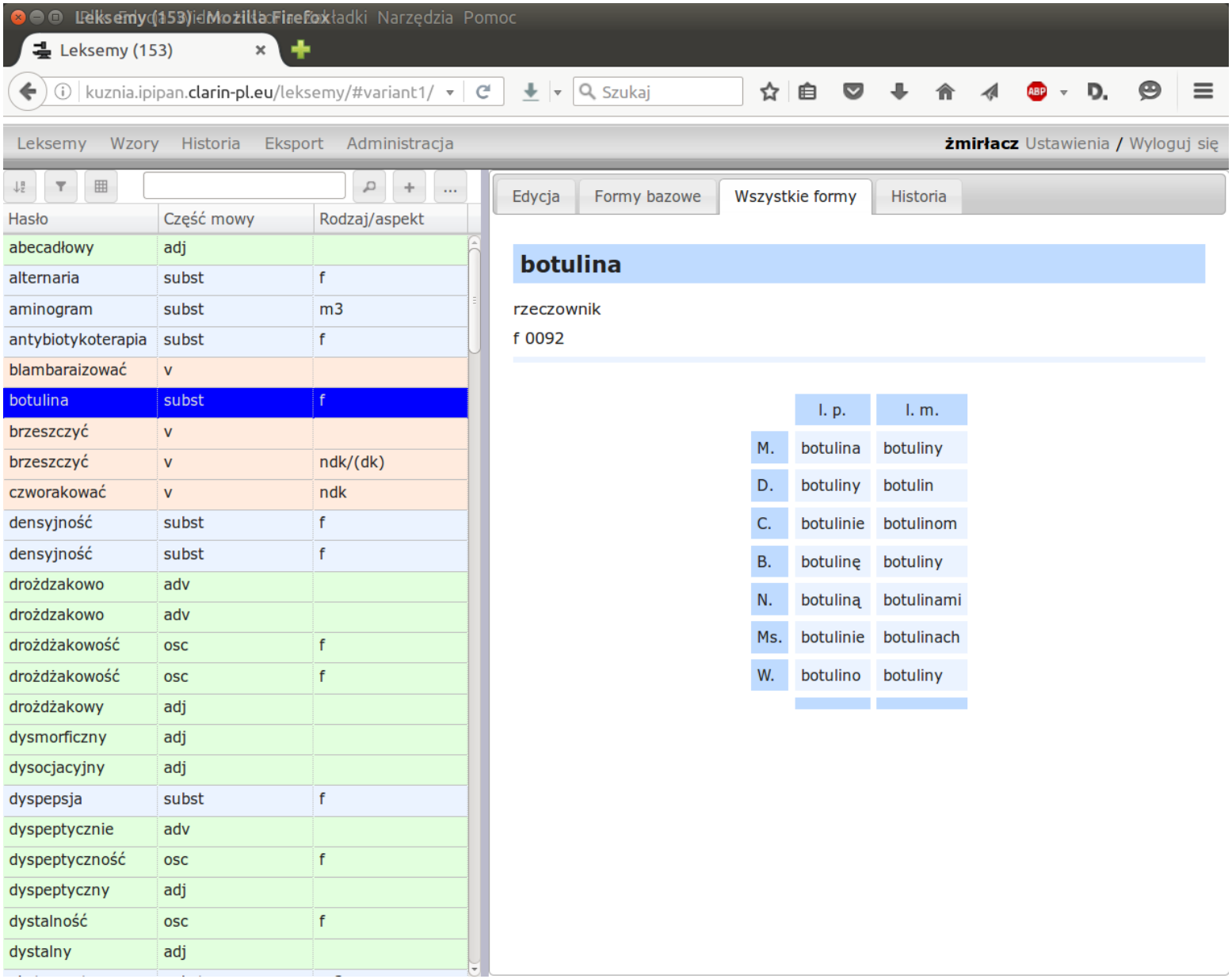
Przykład tworzenia leksemu w Kuźni



Przykład tworzenia leksemu w Kuźni



Przykład tworzenia leksemu w Kuźni



Ujednoznacznianie fleksyjne

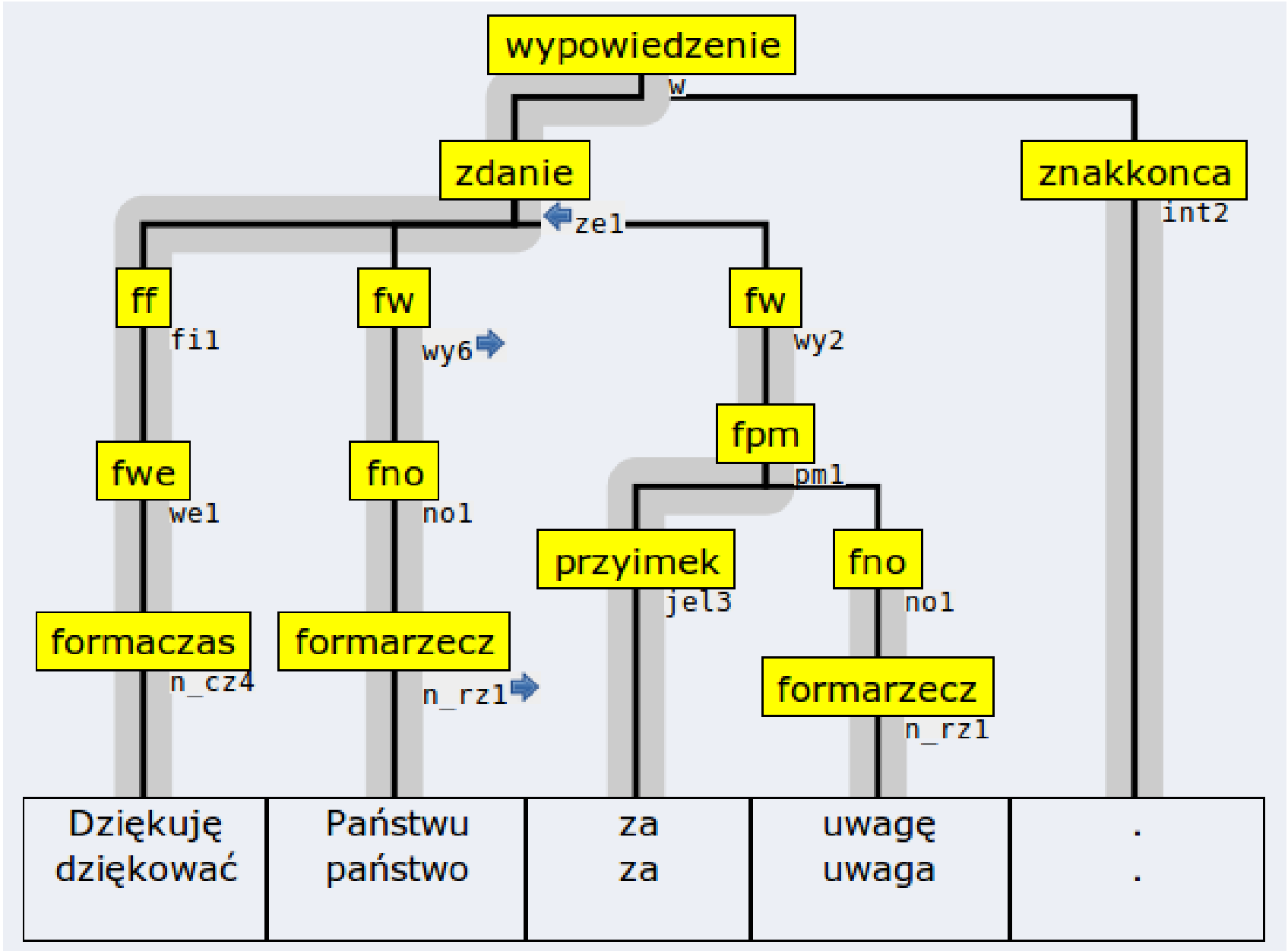
Tager Concraft 2

- program Jakuba Waszczuka:
  `https://github.com/kawu/concraft-pl`
- pracuje bezpośrednio na grafach fleksyjnych Morfeusza (z niejednoznacznościami segmentacji!),
- wytrenowany na milionowym ręcznie znakowanym podkorpusie NKJP,
- zawiera moduł zgadujący znaczniki (ale nie lematy) dla nieznanych słów,
- zawiera również moduł dzielący tekst na zdania.

Tager Concraft 2

Bieżąca wersja sieciowa Morfeusza 2 i Concrafta 2 jest dostępna w Multiserwisie:

`http://multiservice.nlp.ipipan.waw.pl/`


wypowiedzenie
w
zdanie
znakkonca
int2
ze1
ff
fi1
fw
wy6
fw
wy2
fpm
pm1
fwe
we1
fno
no1
przyimek
jel3
fno
no1
formaczas
n_cz4
formarzecz
n_rz1
formarzecz
n_rz1
Dziękuję
dziękować
Państwu
państwo
za
za
uwagę
uwaga
.
.