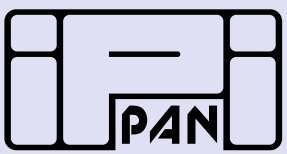
Narzędzia analizy fleksyjnej tekstów polskich

Marcin Woliński i Anna Andrzejczuk



INSTYTUT PODSTAW INFORMATYKI
POLSKIEJ AKADEMII NAUK
ul. Jana Kazimierza 5, 01-248 Warszawa

Warsztaty CLARIN-PL, 15 kwietnia 2015

Wersja 2 programu
opracowana w ramach CLARIN-PL:

`http://sgjp.pl/morfeusz/dopobrania.html`

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

Podstawowe pojęcia

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

Forma (wyrazowa) segment zinterpretowany poprzez przypisanie do leksemu i określenie jego funkcji gramatycznej

Wykładnik (formy) segment reprezentujący ją 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 (wyrazowa) segment zinterpretowany poprzez przypisanie do leksemu i określenie jego funkcji gramatycznej

Wykładnik (formy) segment reprezentujący ją w tekście

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

Technicznie:

Forma trójka ⟨*wykładnik, lemat, znacznik fleksyjny (tag)*⟩

Leksem zbiór form o tym samym lemacie

Podstawowe pojęcia

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

Ujednoznacznianie morfologiczne określenie na podstawie kontekstu, jako którą z możliwych form interpretować dane wystąpienie segmentu

Tagowanie analiza + ujednoznacznienie

Segmentacja form czasu przeszłego

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

Segmentacja form czasu przeszłego

- *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

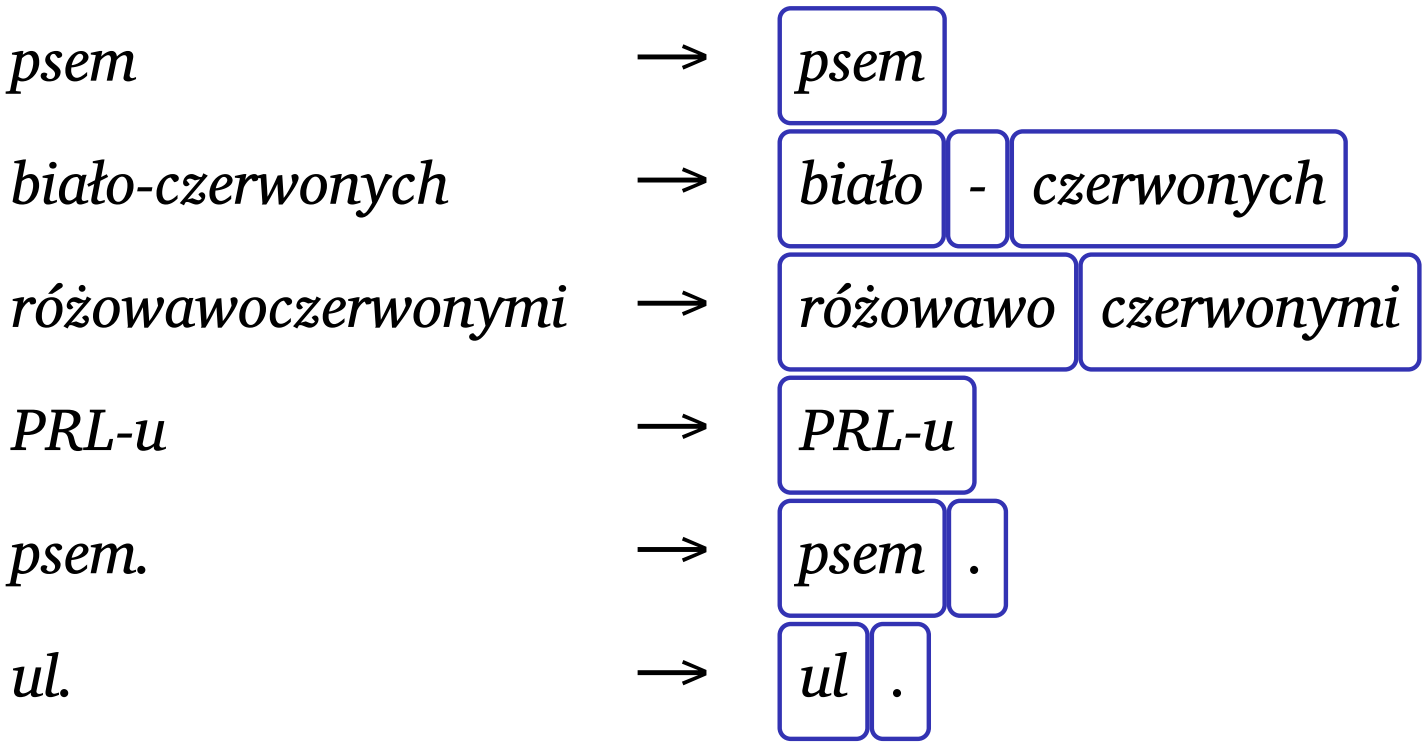
Wariant fundamentalistyczny:

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

Wariant pragmatyczny (nowy znacznik cond):

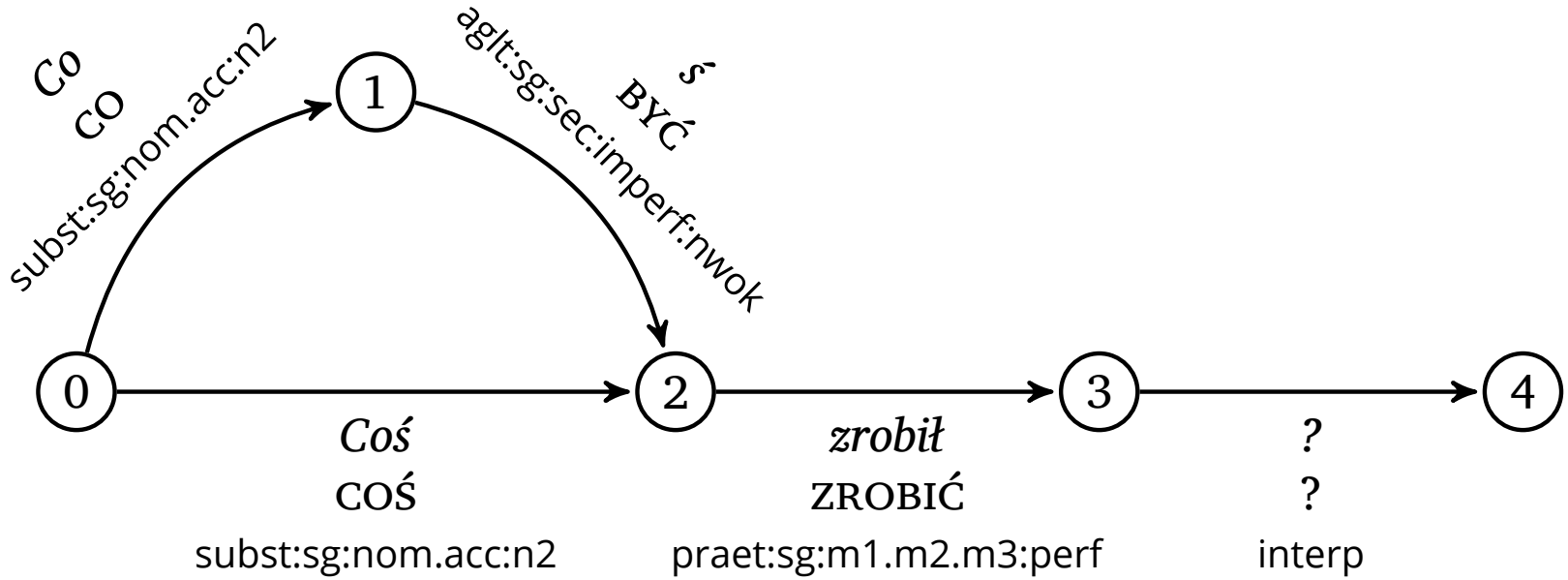
widziałbym	WIDZIEĆ	cond:sg:m1.m2.m3:pri:imperf

Segmentacja



Segmentacja dla polszczyzny jest uwikłana słownikowo.

Reprezentacja wyników analizy



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.

lm

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

bezosobnik: **kra**dziono

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

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
ż	**kra**dła	ś
n	**kra**dło	∅

1.os.
2.os.
3.os.

lm

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

bezosobnik: **kra**dziono

TRYB ROZKAZUJĄCY

lp

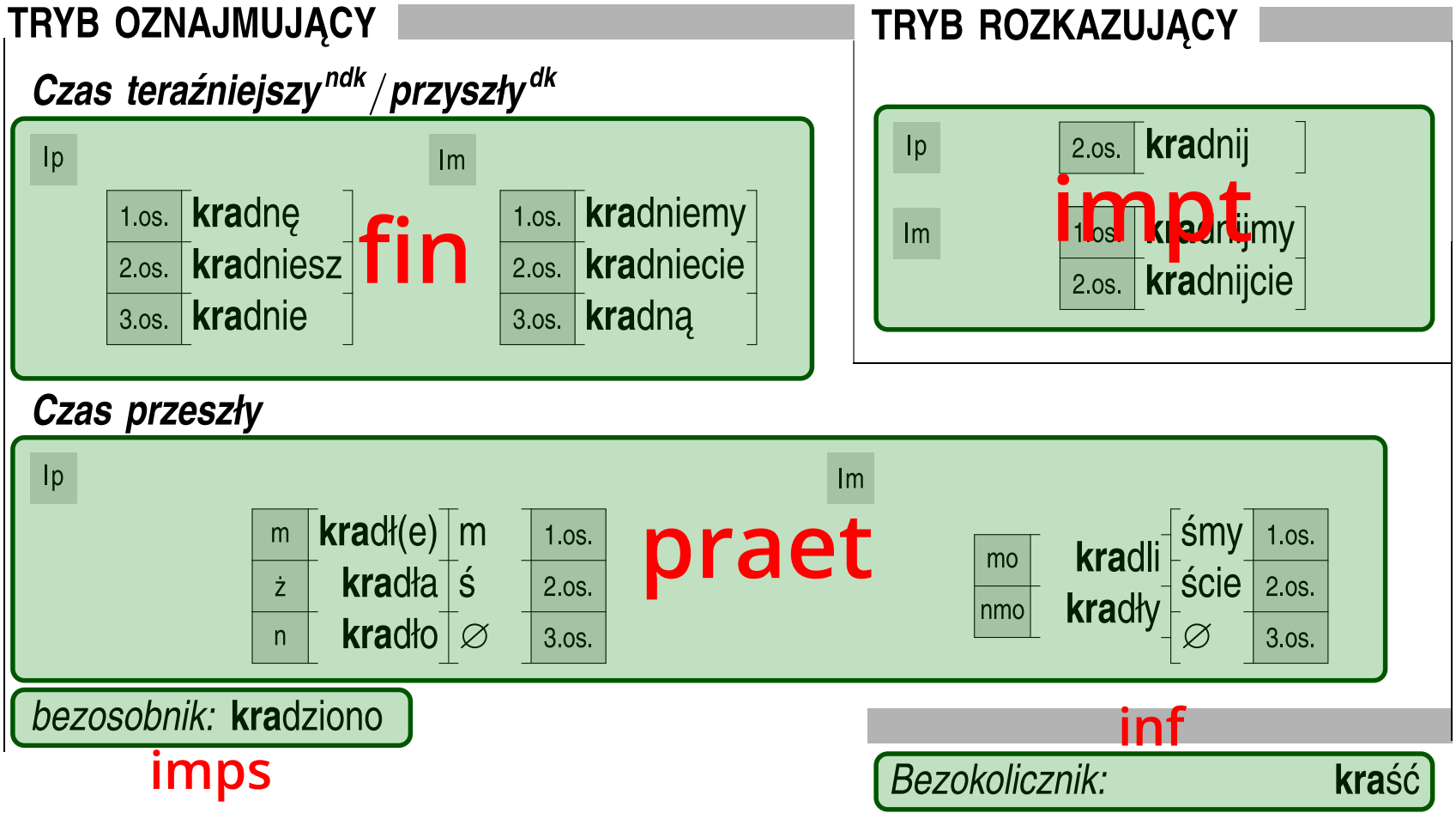
2.os.	**kra**dnij

lm

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.*

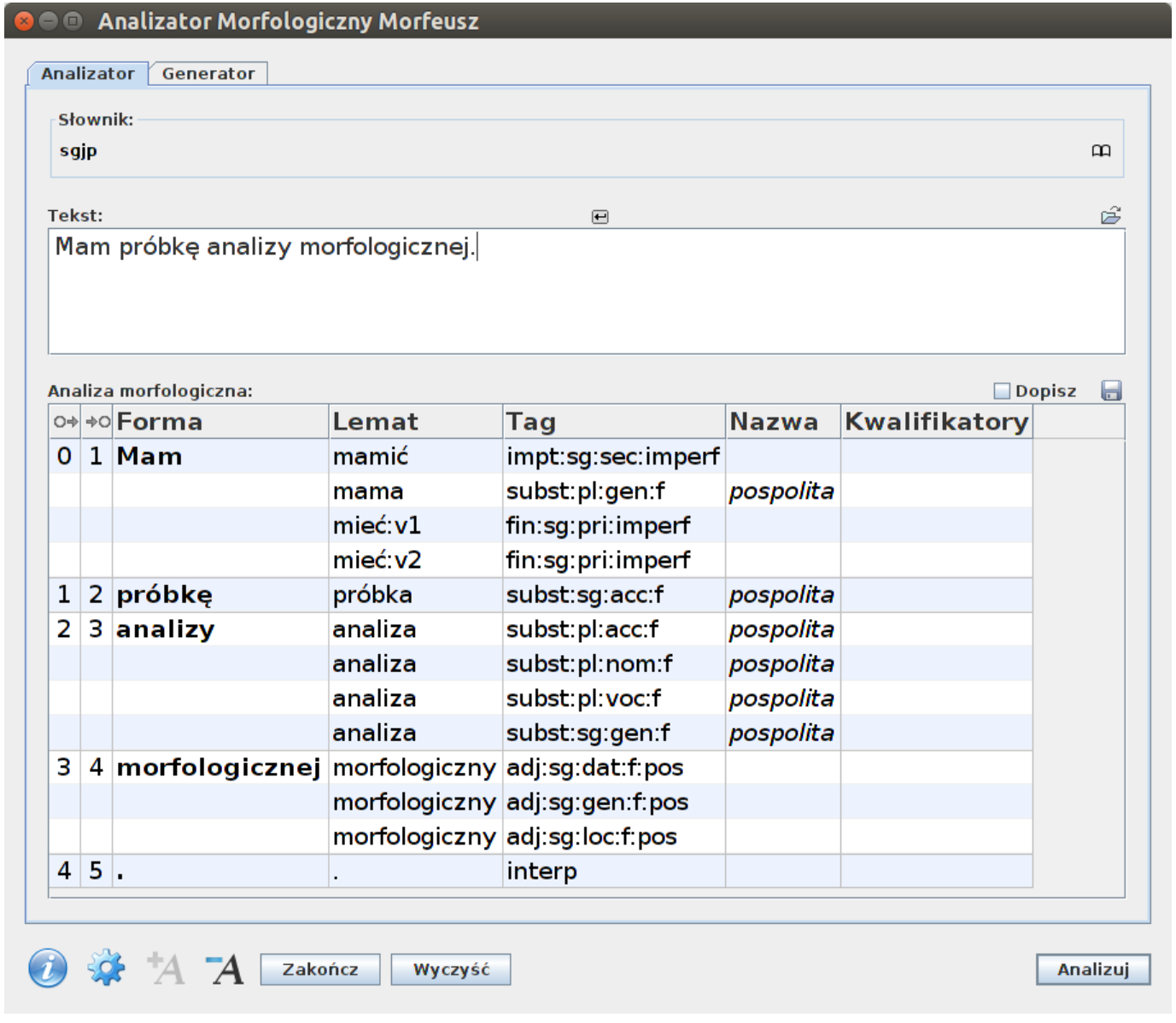
Leksemy ZAMEK:S1 i ZAMEK:S2:

- *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.

Program Morfeusz



		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		

Dystrybucja programu

- Zasadniczą postać programu stanowi moduł programistyczny, który można wbudować w tworzone przez siebie programy.
- 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, Perla, Javy i SWI-Prologu.
- Dla mniej technicznie ukierunkowanych użytkowników przygotowano interfejs graficzny w Javie.

Opcje programu: chętka na aglutynację

- 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, np. `Lem` 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 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.

Opcje programu: słowniki

- Morfeusz jest dystrybuowany z dwoma słownikami: SGJP i Polimorf.
- Kolejne wydania Morfeusza są generowane automatycznie przez system *Kuźnia* zarządzający pracą nad oboma słownikami.
- Aby załadować słownik samemu przygotowany, trzeba najpierw wskazać programowi katalog, w którym miałby takich słowników poszukiwać.

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

Dane wbudowywane w binarny plik słownikowy Morfeusza:

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

Kompilowanie słownika

```
python morfeusz_builder
--input-files=medyczne.tab,sgjp-20150414.tab,dodatki.tab
--tagset-file=morfeusz-sgjp.tagset
--segments-file=segmenty.dat
--dict-dir=morf-dict --dict medyczny
```

(całe powyższe polecenie musi być w jednym wierszu)

Kuźnia

- Do tworzenia słowników w formacie wymaganym przez Morfeusza służy Kuźnia.
- W ramach CLARIN-PL udostępniono instalację Kuźni pozwalającą samodzielnie założyć konto i podjąć pracę nad własnym słownikiem:
  `http://kuznia.ipipan.clarin-pl.eu/accounts/register/`