

Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego

Marcin Woliński



Zespół Inżynierii Lingwistycznej
Instytut Podstaw Informatyki
Polskiej Akademii Nauk

22 lutego 2015

Morfeusz, wersja 2

opracowana w Zespole Inżynierii Lingwistycznej IPI PAN
w ramach CLARIN-PL:

<http://sgjp.pl/morfeusz>

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

Analizator Morfologiczny Morfeusz

Analizator Generator

Słownik:
sgjp

Tekst:
Mam próbkę analizy morfologicznej.

Analiza morfologiczna: Dopisz

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

Zakończ Wyczyść Analizuj

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

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

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:

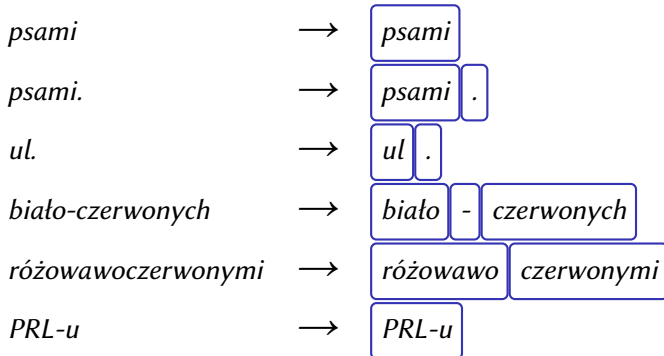
Forma trójka \langle wykładnik, lemat, znacznik fleksyjny (tag) \rangle

Leksem zbiór form o tym samym lemacie

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

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

Tagowanie analiza + ujednoznacznienie



Segmentacja dla polszczyzny jest uwikłana słownikowo.

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

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

Wariant fundamentalistyczny:

<i>widział</i>	WIDZIEĆ	praet:sg:m1.m2.m3:imperf
<i>em</i>	BYĆ	aglt:sg:pri:imperf:wok

Wariant pragmatyczny:

<i>widziałem</i>	WIDZIEĆ	praet:sg:m1.m2.m3:pri:imperf
------------------	---------	------------------------------

Wariant fundamentalistyczny:

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

Wariant pragmatyczny (nowy znacznik cond):

<i>widziałbym</i>	WIDZIEĆ	cond:sg:m1.m2.m3:pri:imperf
-------------------	---------	-----------------------------

TRYB OZNAJMUJĄCY

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

Ip

1.os.	kradnę
2.os.	kradniesz
3.os.	kradnie

Im

1.os.	kradniemy
2.os.	kradniecie
3.os.	kradną

Czas przeszły

Ip

m	kradł(e)	m	1.os.
ż	kradła	ś	2.os.
n	kradło	∅	3.os.

bezosobnik: kradziono

TRYB ROZKAZUJĄCY

Ip

2.os. **kradnij**

Im

1.os. **kradnijmy**
2.os. **kradnijcie**

Im

mo	kradli	śmy	1.os.
nmo		kradły	ście
		∅	3.os.

Bezokolicznik:

kraść

TRYB OZNAJMUJĄCY

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

Ip		Im	
1.os.	kradnę	1.os.	kradniemy
2.os.	kradniesz	2.os.	kradniecie
3.os.	kradnie	3.os.	kradną

TRYB ROZKAZUJĄCY

Ip	2.os.	kradnij
Im	1.os.	kradnijmy
	2.os.	kradnijcie

Czas przeszły

Ip		Im	
m	kradł(e)	m	1.os.
ż	kradła	ś	2.os.
n	kradło	∅	3.os.
mo	kradli	śmy	1.os.
nmo	kradły	ście	2.os.
		∅	3.os.

bezosobnik: kradziono

Bezokolicznik:

kraść

TRYB OZNAJMUJĄCY

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

Ip				Im	
1.os.	kradnę	fin	1.os.	kradniemy	
2.os.	kradniesz		2.os.	kradniecie	
3.os.	kradnie		3.os.	kradną	

TRYB ROZKAZUJĄCY

Ip	2.os.	kradnij
Im	1.os.	kradnijmy
	2.os.	kradnijcie

impt

Czas *przeszły*

Ip				Im		
m	kradł(e)	m	1.os.	praet	śmy	1.os.
ż	kradła	ś	2.os.		ście	2.os.
n	kradło	∅	3.os.		∅	3.os.

bezosobnik: kradziono

imps

inf

Bezokolicznik:

kraść

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

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

Leksem PARA:

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

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.

- 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 (*śle*) i SŁAĆ:V2 (*ściełę*).
- 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.

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

- prostą klasyfikację nazw własnych,
- kwalifikatory.

Wrocławiu, Wrocław, subst:sg:loc.voc:m3, geograficzna, _
Marcina, Marcin, subst:sg:gen.acc:m1, imię, _
tą, ten, adj:sg:inst:f:pos, _, _
tą, ten, adj:sg:acc:f:pos, _, pot.

- Morfeusz jest dystrybuowany z dwoma słownikami: SGJP i Polimorf.
- Kolejne wydania Morfeusza są generowane automatycznie przez system *Kuźnia* zarządzający pracą nad oboma słownikami.

Dane wbudowywane w binarny plik słownikowy Morfeusza:

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

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.
funkjom	funkcja	subst:pl:dat:f	pospolita	
funkcyj	funkcja	subst:pl:gen:f	pospolita	arch.

Kompilator słowników

▶ Stwórz słownik ⚙️ Opcje

Lista plików źródłowych

C:/users/marcin/Moje Dokumenty/IPI/Morfeusz/eksport.tab
C:/users/marcin/Moje Dokumenty/IPI/Morfeusz/sjgp-20150414.tab
C:/users/marcin/Moje Dokumenty/IPI/Morfeusz/dodatki.tab

+
-
↑

Tagset **Katalog docelowy**

Segmenty **Nazwa słownika**

Identyfikator słownika

Informacja o prawach autorskich

- Do tworzenia słowników w formacie wymaganym przez Morfeusza służy system Kuźnia (opracowany w ZIL IPI PAN).
- 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/>