Wielowarstwowy regułowy model fleksji języka polskiego

Wojciech Jaworski, Szymon Rutkowski

Instytut Informatyki Uniwersytetu Warszawskiego

.

Instytut Podstaw Informatyki Polskiej Akademii Nauk

15 października 2018

Model

- Model reprezentuje zasady morfologiczne języka polskiego jako zestaw operacji wykonywanych na obserwowanej formie słowa prowadzących do przekształcenia jej w lemat i zestaw cech morfoskładniowych.
- Celem jest stworzenie reprezentacji polskiej fleksji, która
  - jest zwarta i zrozumiała dla człowieka,
  - odzwierciedla strukturę języka,
  - jest precyzyjna w sposób umożliwiający jej bezpośrednią implementację w postaci *odgadywacza* (ang. *guesser*) oraz generatora form.
- Model został opracowany na podstawie Słownika Gramatycznego Języka Polskiego w wersji z 30.07.2017.

Zakres analizy

- Tworząc model skupiliśmy się na produktywnej części polskiej fleksji, by uchwycić odmianę słów nowych, nieznanych, nie należących do słownika.
- Model nie obejmuje nieregularnych czasowników oraz niewielkiej liczby słów należących do innych części mowy o nieregularnej odmianie.
  - Wynika to stąd, że znany zamknięty zbiór słów można zawrzeć w słowniczku załączonym do modelu.
- Model nie analizuje również form które nie mają widocznych cech fleksyjnych takich jak
  - znaki interpunkcyjne,
  - liczby, daty, itp. zapisane cyframi,
  - skróty.
- Model obejmuje
  - odmianę akronimów,
  - odmianę słów o niepolskiej ortografii,
  - niektóre formy gwarowe.

Niejednoznaczność

- Zadania lematyzacji i anotacji morfosyntaktycznej nie da się wykonać w sposób jednoznaczny jedynie na podstawie obserwacji pojedynczej, wyrwanej z kontekstu formy.
- Guesser określa z pomocą swoich reguł jedynie zbiór możliwych interpretacji.
- Mogą stanowić one dane wejściowe dla *taggera* przeprowadzającego dezambiguację morfosyntaktyczną na podstawie modeli statystycznych.

Warstwy

1. Warstwa **ortograficzno-fonetyczna** abstrahuje od polskiej ortografii przez przekonwertowanie formy segmentu do wewnętrznej reprezentacji, odzwierciedlającej prawidłowości morfonologiczne języka.
2. Warstwa **analityczna** generuje lemat oraz określa występujące afiksy.
3. Warstwa **interpretacji** nadaje segmentowi interpretację morfosyntaktyczną na podstawie wykrytych afiksów.
4. Warstwa korygująca wygenerowane formy i lematy zawierające wygłos.

Warstwa ortograficzno-fonetyczna

- Zadania warstwy ortograficzno-fonetycznej to:
  - wprowadzenie zasady *jeden znak — jeden dźwięk*,
  - wprowadzenie operatora palatalizacji,
  - ujednolicenie ortografii, przykładowo:
    - ★ w słowie „Franz” piszemy przez „z” na końcu, czytamy „c” i odmieniamy tak, jak słowa kończące się na „c”
    - ★ w słowie „ZOZ” piszemy przez „z” na końcu, czytamy „z” i odmieniamy tak, jak słowa kończące się na „z”
    - ★ w słowie „NFZ” piszemy przez „z” na końcu, czytamy „zet” i odmieniamy tak, jak słowa kończące się na „t”
- Konwersja jest odwracalna, ale nie jest jednoznaczna.
- Celem przeprowadzenia tej konwersji jest uproszczenie kolejnych reguł, które mogą korzystać z uogólnień dokonanych już przez tę warstwę.

- W polskim zapisie ortograficznym formy zawierające ten sam rdzeń często różnią się.
- Widać to na przykład w ciągu wyrazów: *pani*, *pań*, *panie*.
- Za pomocą reguł

reguła	prawy kontekst
n′ ← ni	a ą e ę o ó u
n′ ← n	i *sylabotwórcze*
n′ ← ń	*spółgłoska lub wygłos*

  można je przekształcić do postaci: `pan′i`, `pan′`, `pan′e`, gdzie dobrze widoczny jest wspólny rdzeń (`pan′`).
- Domyślna reguła przepisuje znak wejściowy bez zmian; uruchamia się ona, kiedy żadna z innych reguł nie znajduje zastosowania.
- Stosowalność reguł wymaga
  - dopasowania ciągu znaków podlegającego przekształceniu,
  - dopasowania prawego kontekstu (ciągu znaków następującego bezpośrednio po ciągu przekształcanym).

Wybrane reguły ortograficzno-fonetyczne

reguła	reguła	reguła	prawy kontekst
b′ ← bi	p′ ← pi	m′ ← mi	a ą e ę o ó u
b′ ← b	p′ ← p	m′ ← m	i *sylabotwórcze*
v′ ← wi	f′ ← fi		a ą e ę o ó u
v′ ← w	f′ ← f		i *sylabotwórcze*
v ← w		x ← ch	*litera inna niż* i
d′ ← dzi	t′ ← ci	n′ ← ni	a ą e ę o ó u
d′ ← dz	t′ ← c	n′ ← n	i *sylabotwórcze*
d′ ← dź	t′ ← ć	n′ ← ń	*spółgłoska lub wygłos*
z′ ← zi	s′ ← si		a ą e ę o ó u
z′ ← z	s′ ← s		i *sylabotwórcze*
z′ ← ź	s′ ← ś		*spółgłoska lub wygłos*
ǯ ← dż	č ← cz	ʒ ← dz	*litera inna niż* i
ž ← ż	š ← sz	ř ← rz	*litera inna niż* i
g′ ← gi	k′ ← ki		a ą e ę o ó u
g′ ← g	k′ ← k		i *sylabotwórcze*

Podział głosek

- Głoski dzielimy na funkcjonalnie miękkie i funkcjonalnie twarde.
- Funkcjonalnie twarde to takie, które można zmiękczyć, należą do nich:

  b, x, d, f, g, h, k, ł, m, n, p, r, s, t, v, z
- Funkcjonalnie miękkie to takie, których zmiękczyć się nie da, należą do nich:

  b′, t′, d′, f′, m′, n′, p′, s′, v′, z′, l, c, č, ʒ, ǯ, ř, š, ž
- Dalszą analizę będziemy wykonywać osobno dla słów mających funkcjonalnie twardą ostatnią głoskę rdzenia (np. *pan*, *gwiazda*) i tych, które mają ją funkcjonalnie miękką (np. *pani*, *kość*).
- Mają one bowiem różne paradygmaty odmiany ze względu na możliwość wystąpienia sufiksów zmiękczających.

Sufiksy i alternacje

W poniższej tabeli znajdują się wybrane formy rzeczowników *gwiazda*, *sąsiad*, *szpieg* i *waga* oraz przymiotników *rudy* i *nagi* uporządkowane według końcówek.

-a	gwiazda	sąsiada	ruda	szpiega	waga	naga
-om	gwiazdom	sąsiadom		szpiegom	wagom	
-ą	gwiazdą		rudą		wagą	nagą
-e			rude			nagie
-em		sąsiadem		szpiegiem		
-y	gwiazdy	sąsiady	rudy	szpiegi	wagi	nagi
-i		sąsiedzi	rudzi	szpiedzy		nadzy
-ie	gwieździe	sąsiedzie			wadze	
-	gwiazd	sąsiad		szpieg	wag	

Na czerwono zaznaczone są alternacje rdzenia i zmiany końcówki.

Sufiksy i alternacje

Zastosowanie reguł warstwy ortograficzno-fonetycznej upraszcza alternacje.

-a	gv′azda	sąs′ada	ruda	szp′ega	waga	naga
-om	gv′azdom	sąs′adom		szp′egom	wagom	
-ą	gv′azdą		rudą		wagą	nagą
-e			rude			nag′e
-em		sąs′adem		szp′eg′em		
-y	gv′azdy	sąs′ady	rudy	szp′eg′i	wag′i	nag′i
-′i		sąs′ed′i	rud′i	szp′eʒy		naʒy
-′e	gv′ez′d′e	sąs′ed′e			waʒe	
-ε	gv′azd	sąs′ad		szp′eg	wag	

Rodzaje sufiksów

- Ze względu na występujące alternacje końcówki możemy podzielić na
  - neutralne: -a, -ami, -ach, -om, -o, -u
  - zmiękczające głoski g i k: -e, -ego, -ej, -em, -emu
  - -y występujące czasami jako -i: -y, -ych, -ym, -ymi
  - zmiękczające -′i
  - zmiękczające -′e
  - wygłos -ε
- W przypadku głoski g występują dwa rodzaje zmiękczenia: zamiana na g′ oraz zamiana na ʒ.
- Z uwagi na to, że w formie adj:pl:nom:m1:pos w przypadku głoski g następuje zamiana na ʒy, a w przypadku innych głosek mamy tu typową palatalizację, uznajemy ʒy za efekt działania zmiękczającego -′i.
- Analogicznie postępujemy w przypadku paradygmatów rzeczownikowych zmiękczającego -′i oraz zmiękczającego -′e.

Grupy alternacyjne

Dla każdej końcówki możemy wypisać występujące przy nich alternacje:

	αy	αe	α	αi	αie	$\alpha\varepsilon$
d	dy $\rightarrow$ d	de $\rightarrow$ d	d $\rightarrow$ d	ed′i $\rightarrow$ ad	ed′e $\rightarrow$ ad ez′d′e $\rightarrow$ azd	d $\rightarrow$ d
g	g′i $\rightarrow$ g	g′e $\rightarrow$ g	g $\rightarrow$ g	ʒy $\rightarrow$ g	ʒe $\rightarrow$ g	g $\rightarrow$ g

- W nagłówku tabeli umieszczone są nazwy grup alternacji.
- Nazwy składają się z
  - symbolu α oznaczającego głoskę funkcjonalnie twardą oraz
  - jednej lub dwu liter oznaczających sufiks.
- Zaznaczone w nazwie litery sufiksu są włączone do alternacji.

Tabela alternacji dla głosek funkcjonalnie twardych

	αy	αe	α	αi	αie	$\alpha\varepsilon$
x	xy → x	xe → x	x → x	s′i → x	še → x	x → x ex → x
d	dy → d	de → d	d → d	d′i → d ed′i → ad	d′e → d z′d′e → zd ed′e → ad ed′e → od ez′d′e → azd	d → d ed → d ód → od ąd → ęd
f	fy → f	fe → f	f → f	f′i → f	f′e → f	f → f
h	hy → h	he → h	h → h	z′i → h	še → h že → h	h → h
m	my → m	me → m	m → m	m′i → m s′m′i → sm	m′e → m s′m′e → sm	m → m em → m
r	ry → r	re → r	r → r	ry → r	re → r eře → ar etře → atr ře → rr	r → r er → r ′er → r ór → or ⋆cer → kr óbr → obr óstr → ostr
k	k′i → k	k′e → k	k → k	cy → k	ce → k	k → k ek → k ąk → ęk
...	...					

Reguły analityczne

- Możemy teraz zdefiniować reguły opisujące zmiany następujące podczas dodawania sufiksu do rdzenia.
- Każda reguła składa się z opisu modyfikacji wykonywanych na przetwarzanej formie oraz zbioru definiowanych atrybutów.
- Reguły te są parametryzowane przez grupy alternacyjne.
- Przykładowa reguła ucinająca końcówkę „ego” u przymiotników:

$$-\alpha\mathrm{e}\ \mathrm{go} \quad \text{flex:=ego palat:=n cat:=adj}$$

- Po zastąpieniu $\alpha\mathrm{e}$ przez kolejne alternacje należące do tej grupy otrzymujemy reguły

$$\mathrm{dego} \to \mathrm{d} \quad \text{flex:=ego palat:=n con:=d cat:=adj}$$
$$\mathrm{g'ego} \to \mathrm{g} \quad \text{flex:=ego palat:=n con:=g cat:=adj}$$

  zamieniające *rudego* na *rud* oraz *nag' ego* na *nag*.
- Wartość atrybutu con jest dodawana podczas rozwijania reguły na podstawie wybranej alternacji.

Reguły ucinające sufiks formy i dodające sufiks lematu rzeczownika z głoską funkcjonalnie twardą

$$
\begin{bmatrix}
-\alpha\text{y} & \text{flex} := \text{y}, \downarrow, \text{noun} \\
-\alpha\text{e} & \text{flex} := \text{e}, \downarrow, \text{noun} \\
-\alpha\text{e m} & \text{flex} := \text{em}, \downarrow, \text{noun} \\
-\alpha\text{ a} & \text{flex} := \text{a}, \downarrow, \text{noun} \\
-\alpha\text{ ax} & \text{flex} := \text{ach}, \downarrow, \text{noun} \\
-\alpha\text{ am}'\text{i} & \text{flex} := \text{ami}, \downarrow, \text{noun} \\
-\alpha\text{ ą} & \text{flex} := \text{ą}, \downarrow, \text{noun} \\
-\alpha\text{ ę} & \text{flex} := \text{ę}, \downarrow, \text{noun} \\
-\alpha\text{ o} & \text{flex} := \text{o}, \downarrow, \text{noun} \\
-\alpha\text{ om} & \text{flex} := \text{om}, \downarrow, \text{noun} \\
-\alpha\text{ ov}'\text{i} & \text{flex} := \text{owi}, \downarrow, \text{noun} \\
-\alpha\text{ ov}'\text{e} & \text{flex} := \text{owie}, \downarrow, \text{noun} \\
-\alpha\text{ óv} & \text{flex} := \text{ów}, \downarrow, \text{noun} \\
-\alpha\text{ u} & \text{flex} := \text{u}, \downarrow, \text{noun} \\
-\alpha\text{ um} & \text{flex} := \text{um}, \downarrow, \text{noun} \\
-\alpha\text{i} & \text{flex} := \text{i}, \downarrow, \text{noun} \\
-\alpha\text{ie} & \text{flex} := \text{ie}, \downarrow, \text{noun} \\
-\alpha\varepsilon & \text{flex} := \varepsilon, \downarrow, \text{noun} \\
\star - \alpha\varepsilon\text{ m}'\text{i} & \text{flex} := \text{ami}, \downarrow, \text{noun}
\end{bmatrix}
\otimes
\begin{bmatrix}
+\alpha\text{y} & \text{lemma} := \text{y} \\
+\alpha\text{e} & \text{lemma} := \text{e} \\
+\alpha\text{ a} & \text{lemma} := \text{a} \\
+\alpha\text{ o} & \text{lemma} := \text{o} \\
+\alpha\text{ ov}'\text{e} & \text{lemma} := \text{owie} \\
+\alpha\text{ um} & \text{lemma} := \text{um} \\
\star + \alpha\text{ us} & \text{lemma} := \text{us} \\
+\alpha\text{i} & \text{lemma} := \text{i} \\
+\alpha\varepsilon & \text{lemma} := \varepsilon
\end{bmatrix}
$$

Reguły dla końcówek ych, ym, ymi, ego, ej, emu zostały pominięte.
Symbol $+$ oznacza, że reguła przykleja sufiks.

Rozpakowywanie reguł

Rozpatrzmy alternacje

	αy	αe	α	αi	αie	$\alpha\varepsilon$
d	dy $\rightarrow$ d	de $\rightarrow$ d	d $\rightarrow$ d	ed′i $\rightarrow$ ad	ed′e $\rightarrow$ ad ez′d′e $\rightarrow$ azd	d $\rightarrow$ d
g	g′i $\rightarrow$ g	g′e $\rightarrow$ g	g $\rightarrow$ g	ʒy $\rightarrow$ g	ʒe $\rightarrow$ g	g $\rightarrow$ g

oraz reguły analityczne

$$\left[\begin{array}{ll} -\alpha\ \mathrm{a} & \mathrm{flex} := \mathrm{a}, \downarrow, \mathrm{noun} \\ -\alpha\mathrm{e}\ \mathrm{m} & \mathrm{flex} := \mathrm{em}, \downarrow, \mathrm{noun} \\ -\alpha\mathrm{ie} & \mathrm{flex} := \mathrm{ie}, \downarrow, \mathrm{noun} \end{array}\right] \otimes \left[\begin{array}{ll} +\alpha\ \mathrm{a} & \mathrm{lemma} := \mathrm{a} \\ +\alpha\varepsilon & \mathrm{lemma} := \varepsilon \end{array}\right]$$

Po rozwinięciu alternacji otrzymamy reguły:

$$\left[\begin{array}{ll} \mathrm{da} \rightarrow \mathrm{d} & \mathrm{flex} := \mathrm{a}, \downarrow, \mathrm{con} := \mathrm{d}, \mathrm{noun} \\ \mathrm{ga} \rightarrow \mathrm{g} & \mathrm{flex} := \mathrm{a}, \downarrow, \mathrm{con} := \mathrm{g}, \mathrm{noun} \\ \mathrm{dem} \rightarrow \mathrm{d} & \mathrm{flex} := \mathrm{em}, \downarrow, \mathrm{con} := \mathrm{d}, \mathrm{noun} \\ \mathrm{g'em} \rightarrow \mathrm{g} & \mathrm{flex} := \mathrm{em}, \downarrow, \mathrm{con} := \mathrm{g}, \mathrm{noun} \\ \mathrm{ed'e} \rightarrow \mathrm{ad} & \mathrm{flex} := \mathrm{ie}, \downarrow, \mathrm{con} := \mathrm{d}, \mathrm{noun} \\ \mathrm{ez'd'e} \rightarrow \mathrm{azd} & \mathrm{flex} := \mathrm{ie}, \downarrow, \mathrm{con} := \mathrm{d}, \mathrm{noun} \\ \mathrm{ʒe} \rightarrow \mathrm{g} & \mathrm{flex} := \mathrm{ie}, \downarrow, \mathrm{con} := \mathrm{g}, \mathrm{noun} \end{array}\right] \otimes \left[\begin{array}{ll} \mathrm{d} \rightarrow \mathrm{da} & \mathrm{lemma} := \mathrm{a} \\ \mathrm{g} \rightarrow \mathrm{ga} & \mathrm{lemma} := \mathrm{a} \\ \mathrm{d} \rightarrow \mathrm{d} & \mathrm{lemma} := \varepsilon \\ \mathrm{g} \rightarrow \mathrm{g} & \mathrm{lemma} := \varepsilon \end{array}\right]$$

Rozpakowywanie reguł cd.

$$\begin{bmatrix} da \to d & flex := a, \downarrow, con := d, noun \\ ga \to g & flex := a, \downarrow, con := g, noun \\ dem \to d & flex := em, \downarrow, con := d, noun \\ g'em \to g & flex := em, \downarrow, con := g, noun \\ ed'e \to ad & flex := ie, \downarrow, con := d, noun \\ ez'd'e \to azd & flex := ie, \downarrow, con := d, noun \\ ʒe \to g & flex := ie, \downarrow, con := g, noun \end{bmatrix} \otimes \begin{bmatrix} d \to da & lemma := a \\ g \to ga & lemma := a \\ d \to d & lemma := \varepsilon \\ g \to g & lemma := \varepsilon \end{bmatrix}$$

Teraz łączymy reguły z pierwszej kolumny z tymi z kolumny drugiej:

$$\begin{array}{ll} da \to da & flex := a, \downarrow, con := d, lemma := a, noun \\ ga \to ga & flex := a, \downarrow, con := g, lemma := a, noun \\ dem \to da & flex := em, \downarrow, con := d, lemma := a, noun \\ g'em \to ga & flex := em, \downarrow, con := g, lemma := a, noun \\ ed'e \to ada & flex := ie, \downarrow, con := d, lemma := a, noun \\ ez'd'e \to azda & flex := ie, \downarrow, con := d, lemma := a, noun \\ ʒe \to ga & flex := ie, \downarrow, con := g, lemma := a, noun \\ da \to d & flex := a, \downarrow, con := d, lemma := \varepsilon, noun \\ ga \to g & flex := a, \downarrow, con := g, lemma := \varepsilon, noun \\ dem \to d & flex := em, \downarrow, con := d, lemma := \varepsilon, noun \\ g'em \to g & flex := em, \downarrow, con := g, lemma := \varepsilon, noun \\ ed'e \to ad & flex := ie, \downarrow, con := d, lemma := \varepsilon, noun \\ ez'd'e \to azd & flex := ie, \downarrow, con := d, lemma := \varepsilon, noun \\ ʒe \to g & flex := ie, \downarrow, con := g, lemma := \varepsilon, noun \end{array}$$

Rozpakowywanie reguł cd.

da → da	flex := a, ↓, con := d, lemma := a, noun
ga → ga	flex := a, ↓, con := g, lemma := a, noun
dem → da	flex := em, ↓, con := d, lemma := a, noun
g′em → ga	flex := em, ↓, con := g, lemma := a, noun
ed′e → ada	flex := ie, ↓, con := d, lemma := a, noun
ez′d′e → azda	flex := ie, ↓, con := d, lemma := a, noun
ʒe → ga	flex := ie, ↓, con := g, lemma := a, noun
da → d	flex := a, ↓, con := d, lemma := ε, noun
ga → g	flex := a, ↓, con := g, lemma := ε, noun
dem → d	flex := em, ↓, con := d, lemma := ε, noun
g′em → g	flex := em, ↓, con := g, lemma := ε, noun
ed′e → ad	flex := ie, ↓, con := d, lemma := ε, noun
ez′d′e → azd	flex := ie, ↓, con := d, lemma := ε, noun
ʒe → g	flex := ie, ↓, con := g, lemma := ε, noun

Rozpakowane reguły możemy użyć do lematyzacji form:

gv′azda → gv′azda	flex := a, ↓, con := d, lemma := a, noun
gv′azda → gv′azd	flex := a, ↓, con := d, lemma := ε, noun
szp′eg′em → szp′ega	flex := em, ↓, con := g, lemma := a, noun
szp′eg′em → szp′eg	flex := em, ↓, con := g, lemma := ε, noun
gv′ez′d′e → gv′azda	flex := ie, ↓, con := d, lemma := a, noun
gv′ez′d′e → gv′azd	flex := ie, ↓, con := d, lemma := ε, noun
waʒe → waga	flex := ie, ↓, con := g, lemma := a, noun
waʒe → wag	flex := ie, ↓, con := g, lemma := ε, noun

Warstwa interpretacji

- Warstwa interpretacji zawiera reguły przypisujące interpretację morfosyntaktyczną na podstawie wartości atrybutów.
- Warstwa ta dokonuje selekcji kandydatów powstałych w wyniku działania warstwy analitycznej (wprowadzając jednocześnie kolejną niejednoznaczność).

$$
\begin{array}{lcl}
\text{flex} := \text{a}, \downarrow, \text{lemma} := \text{a}, \text{noun} & \rightarrow & \text{subst:sg:nom:m1.m2.f} \\
\text{flex} := \text{a}, \downarrow, \text{lemma} := \text{a}, \text{noun} & \rightarrow & \text{subst:pl:nom.acc.voc:n:pt} \\
\text{flex} := \text{a}, \downarrow, \text{lemma} := \varepsilon, \text{noun} & \rightarrow & \text{subst:sg:gen.acc:m1.m2} \\
\text{flex} := \text{a}, \downarrow, \text{lemma} := \varepsilon, \text{noun} & \rightarrow & \text{subst:sg:gen:m3} \\
\text{flex} := \text{em}, \downarrow, \text{lemma} := \varepsilon, \text{noun} & \rightarrow & \text{subst:sg:inst:m1.m2.m3}
\end{array}
$$

- Dla rzeczowników jest to najmniej ustrukturalizowana warstwa.
- W przypadku czasowników, przymiotników i przysłówków to odwzorowanie jest dość jednoznaczne.
- Reguły interpretacji zostały wytworzone półautomatycznie na podstawie SGJP.

Działanie warstwy interpretacji
Reguły interpretacji

$$
\begin{array}{lcl}
\text{flex} := \text{a}, \downarrow, \text{lemma} := \text{a}, \text{noun} & \rightarrow & \text{subst:sg:nom:m1.m2.f} \\
\text{flex} := \text{a}, \downarrow, \text{lemma} := \text{a}, \text{noun} & \rightarrow & \text{subst:pl:nom.acc.voc:n:pt} \\
\text{flex} := \text{a}, \downarrow, \text{lemma} := \varepsilon, \text{noun} & \rightarrow & \text{subst:sg:gen.acc:m1.m2} \\
\text{flex} := \text{a}, \downarrow, \text{lemma} := \varepsilon, \text{noun} & \rightarrow & \text{subst:sg:gen:m3} \\
\text{flex} := \text{em}, \downarrow, \text{lemma} := \varepsilon, \text{noun} & \rightarrow & \text{subst:sg:inst:m1.m2.m3}
\end{array}
$$

przypiszą formom

$$
\begin{array}{ll}
\text{gv}'\text{azda} \rightarrow \text{gv}'\text{azda} & \text{flex} := \text{a}, \downarrow, \text{con} := \text{d}, \text{lemma} := \text{a}, \text{noun} \\
\text{gv}'\text{azda} \rightarrow \text{gv}'\text{azd} & \text{flex} := \text{a}, \downarrow, \text{con} := \text{d}, \text{lemma} := \varepsilon, \text{noun} \\
\text{szp}'\text{eg}'\text{em} \rightarrow \text{szp}'\text{ega} & \text{flex} := \text{em}, \downarrow, \text{con} := \text{g}, \text{lemma} := \text{a}, \text{noun} \\
\text{szp}'\text{eg}'\text{em} \rightarrow \text{szp}'\text{eg} & \text{flex} := \text{em}, \downarrow, \text{con} := \text{g}, \text{lemma} := \varepsilon, \text{noun}
\end{array}
$$

następujące interpretacje morfosyntaktyczne:

$$
\begin{array}{ll}
\text{gv}'\text{azda} \rightarrow \text{gv}'\text{azda} & \text{subst:sg:nom:m1.m2.f} \\
\text{gv}'\text{azda} \rightarrow \text{gv}'\text{azda} & \text{subst:pl:nom.acc.voc:n:pt} \\
\text{gv}'\text{azda} \rightarrow \text{gv}'\text{azd} & \text{subst:sg:gen.acc:m1.m2} \\
\text{gv}'\text{azda} \rightarrow \text{gv}'\text{azd} & \text{subst:sg:gen:m3} \\
\text{szp}'\text{eg}'\text{em} \rightarrow \text{szp}'\text{eg} & \text{subst:sg:inst:m1.m2.m3}
\end{array}
$$

Quasi-paradygmaty odmiany

- Reguły przypisujące interpretacje można pogrupować ze względu na wartość atrybutu lemma i rodzaj rzeczownika generowany przez regułę.
- Uzyskujemy w ten sposób „quasi-paradygmaty” odmiany rzeczowników.
- Należy jednak pamiętać, że dany lemat nie jest do takich „paradygmatów” sztywno przypisany:
  - nie musi on mieć form pochodzących tylko z jednego paradygmatu i
  - nie musi mieć wszystkich form występujących w danym paradygmacie.

Rzeczowniki z wygłosem w lemacie

cat=noun lemma=ε gender:=f

sg:nom.acc	ε ↑
sg:gen.dat.loc.voc|pl:gen	y↑
sg:inst	ą↑
pl:nom.acc.voc	y↑ e↑
pl:dat	om↑
pl:inst	ami↑
pl:loc	ach↑

cat=noun lemma=ε gender:=m1

sg:nom	ε
sg:gen	⋆y
sg:gen.acc	a
sg:dat	owi ⋆u
sg:dat.loc	⋆y
sg:acc	⋆y
sg:inst	em ⋆ą
sg:loc	⋆u ⋆ie
sg:loc.voc	u↑ u↓→ ie↓←
sg:voc	cze↑ ⋆y ⋆ie
pl:nom.voc	y↑ i↓ e↑ owie ⋆ie
pl:gen.acc	ów y↑
pl:dat	om
pl:inst	ami
pl:loc	ach
depr	y↓ e↑

cat=noun lemma=ε gender:=m2

sg:nom	ε
sg:gen.acc	a
sg:dat	owi ⋆u
sg:inst	em
sg:loc.voc	u↑ u↓→ ie↓←
pl:nom.acc.voc	y↓ e↑ ⋆e
pl:gen	ów y↑
pl:dat	om
pl:inst	ami
pl:loc	ach

cat=noun lemma=ε gender:=m3

sg:nom.acc	ε
sg:gen	u a
sg:dat	⋆u↓ owi
sg:inst	em
sg:loc	⋆ie
sg:loc.voc	u↑ u↓→ ie↓←
sg:voc	⋆ie
pl:nom.acc.voc	y↓ e↑ ⋆e ⋆a
pl:gen	ów y↑
pl:dat	om
pl:inst	ami
pl:loc	ach

Rzeczowniki z kończące się na „a” w lemacie

cat=noun lemma=a gender:=f

sg:nom	a
sg:gen	y ⋆ε
sg:gen.dat.loc	ej
sg:dat.loc	y↑ ie↓
sg:acc	ę ą
sg:inst	ą
sg:voc	u↑ o a
pl:nom.acc.voc	y↓ e
pl:gen	ε y↑
pl:gen.loc	ych
pl:dat	ym om
pl:inst	ymi ami
pl:loc	ach

cat=noun lemma=a gender:=m1

sg:nom	a
sg:gen	y ⋆ego
sg:gen.acc	⋆ego
sg:dat	⋆emu
sg:dat.loc	y↑ ie↓
sg:acc	ę
sg:inst	ą ⋆ym
sg:loc	⋆ym
sg:voc	o ⋆u
pl:nom.voc	y↑ i↓ owie ⋆e
pl:gen.acc	ów ⋆ε
pl:dat	om
pl:inst	ami
pl:loc	ach
depr	y e

Lista nieobecności

- Głoski funcjonalnie miękkie.
- Leksemy typu „-cja”, „-pia”, „-dia”, „-rium”.
- Słowa pisane z użyciem obcej ortografii.
- Odmiana akronimów
- Odmiana (stopniowanie) przymiotników i przysłówków
- Odmiana czasowników
- Postprocessing wygłosu
- Formy gwarowe

Reguły operacyjne

- Model składa się z
  - 723 reguł warstwy ortograficzno-fonetycznej
  - 748 alternacji
  - 367 reguł analitycznych
  - 960 reguł przypisujących interpretację
- W celu wytworzenia wydajnego systemu reguły te zostały złączone ze sobą:
  - do każdej możliwej sekwencji reguł analitycznych
  - zostały dopasowane reguły przypisujące interpretację;
  - następnie zostały przekonwertowane na standardową ortografię.
- W wyniku tego procesu powstało ok. 10 000 000 reguł operacyjnych.
- Następnie została dokonana selekcja reguł polegająca na wyborze tych, których użycie jest poświadczone w SGJP uzupełnionym o przykładowe formy gwarowe i dodatkowe odmienione akronimy.
- Reguł operacyjnych jest 31122.

Reguły operacyjne

Liczbę reguł z podziałem na ich typy i części mowy:

	noun	adj	adv	verb	suma
produktywne	7534	1501	150	9107	18292
* nieproduktywne	209	389	—	3701	4299
A obce	1275	—	—	—	1275
B obce	206	—	—	—	206
C akronimy	557	—	—	—	557
D gwarowe	2639	380	—	3474	6493
suma	12420	2270	150	16282	31122

- Grupa „obcych A” dotyczy słów o obcej ortografii, w których pierwotna postać rdzenia jest zawarta w obserwowanej formie.
- W wypadku „obcych B” pierwotna postać rdzenia nie jest zawarta w obserwowanej formie i musi zostać odgadnięta (np. dopełniacz *Chiraka* od lematu *Chirac*).
- Wszystkim regułom towarzyszą informacje o frekwencji — liczba form ze słownika lematyzowalnych za pomocą danej reguły.

Pokrycie modelu

- Reguły produkcyjne opisują fleksję
  - $\frac{143643}{143643+343}$ = 99,76% lematów rzeczownikowych,
  - $\frac{66426}{66426+26}$ = 99,96% lematów przymiotnikowych,
  - $\frac{25839}{25839+422}$ = 98,39% lematów przysłówkowych,
  - $\frac{28571}{28571+1229}$ = 95,88% lematów czasownikowych.
- Po usunięciu lematów czasownikowych, które powstały poprzez dodanie prefiksu wartość wzrasta do $\frac{13852}{13852+167}$ = 98,81%.
- Takie wartości wskazują, że opisywany model w sposób poprawny i pełny opisuje zawartą w SGJP fleksję języka polskiego.
- Leksemy niepokryte przez model odmieniają się w sposób nieregularny – powinny one stanowić zamknięty zbiór.
- Jest to szczególnie istotne przy czasownikach, gdzie 167 nieregularnych rdzeni generuje, po uzupełnieniu o prefiksy, 1229 nieregularnych leksemów.
- W przypadku przysłówków, na 422 niepokryte przez model leksemy składają się zasadniczo przysłówki niestopniowalne i niepochodzące od przymiotników.

Dezambiguacja

- Rezultaty zwracane przez model są zazwyczaj wysoce niejednoznaczne.
- W celu ich ujednoznacznienia można podjąć następujące kroki
  - konfrontacja wyniku z SGJP
  - weryfikacja za pomocą listy znanych lematów
  - dezambiguacja statystyczna wykonywana przez tager

Konfrontacja z SGJP

- Reguły produkcyjne mają swoje identyfikatory.
- Na podstawie SGJP została wytworzona lista rdzeni wraz przypisanymi im identyfikatorami reguł właściwych dla danego rdzenia.
- Interpretacje potwierdzone przez listę zostają opatrzone statusem „LemmaVal”.
- Formy z SJGP niepokryte przez model zostały umieszczone w osobnym słowniczku.
- Interpretacje uzyskane za pomocą tego słowniczka są opatrzone statusem „LemmaAlt”.
- Pozostałe interpretacje są oznaczone jako „LemmNotVal”.
- Jeśli odgadywacz nie znajdzie żadnej interpretacji dla danej formy zwracają ze statusem „TokNotFound”.
- W ten sposób odgadywacz uzyskuje pełne pokrycie na SGJP i funkcjonalność analizatora morfologicznego.

Bazy form podstawowych słów

- SGJP
  - ponad 333000 lematów
- SAWA
- TERYT
  - 304 powiaty
  - 38889 miejscowości
  - 24508 części miejscowości
  - 42871 ulice (11272 z nich mają osobowego patrona)
- nazwiska-polskie.pl
  - ponad 220000 nazwisk
- Wikipedia/DBpedia
- Geonames - nazwy geograficzne
- KRS - nazwy organizacji

Algorytm dezambiguacji symbolicznej

- Algorytm polega na przypisaniu interpretacjom priorytetów i wyborze tych interpretacji, które mają najmniejszy priorytet.
- Kryteria wyboru priorytetu:

1	lemat jest na liście znanych lematów	+						-	
2	lemat jest w SGJP	+			-				
3	lematyzacja przeprowadzona zgodnie z SGJP	+	-					+	-
4	tag „no-sgjp”		+	-					
5	forma nieodmienna				+		-		
6	tag „poss-ndm”				+	-			
	priorytet	1	1	R	1	R	1	2	R

- Interpretacje z priorytetem oznaczonym „R” są odrzucane, gdy spełniony jest przynajmniej jeden z warunków:
  - forma została wydzielona z tekstu przy z odciętym aglutynatem,
  - forma została zlematyzowana ze zmienioną wielkością liter,
  - forma została zlematyzowana za pomocą reguły typu **B**.
- Jeśli interpretacja z priorytetem oznaczonym „R” nie zostaje odrzucona otrzymuje priorytet 3.

Struktura form słownych w NKJP1M

	Liczba unikalnych form	Liczba form	Procent unikalnych form	Procent form
lematyzowane przez SGJP	156565	906513	85,4720%	74,6117%
symbole	5796	250926	3,1642%	20,6528%
poprawne spoza SGJP	16581	42195	9,0519%	3,4729%
formy z dywizem i apostrofem	659	783	0,3598%	0,0644%
pt lematyzowane do sg przez SGJP	168	461	0,0917%	0,0379%
tag inny niż proponowany przez SGJP	1151	11020	0,6284%	0,9070%
formy gwarowe bądź archaiczne	132	166	0,0721%	0,0137%
powszechny błąd	156	393	0,0852%	0,0323%
zapis fonetyczny	166	191	0,0906%	0,0157%
literówka	1415	1728	0,7725%	0,1422%
niepoprawny tag	383	593	0,2091%	0,0488%
błąd tokenizacji	5	5	0,0027%	0,0004%
cały korpus	183177	1214974	100,0000%	100,0000%

- Pierwsze dwie kategorie oraz ostatnie pięć nie stanowi ciekawych danych do testowania odgadywacza.
- Pozostałe pięć kategorii wykorzystaliśmy do przeprowadzenia walidacji.

Walidacja

- Odgadywacz został porównany z następującymi programami:
  - Analizator morfologiczny SAM (1996)
  - TaKIPI (2007)
- Wygrywa to porównanie niejako walkowerem z uwagi na to że:
  - SAM korzysta z innego tagsetu niż wszystkie obecne narzędzia (m.in. nie rozróżnia fleksemów form czasownika i segmentów nieodmiennych.
  - TaKIPI wymaga Morfeusza w wersji SIaT (rzuca wyjątek, gdy biblioteka libmorfeusz zwróci tag morfosyntaktyczny comp, interj, brev lub burk).
  - SAM generuje segmentation fault dla niektórych segmentów, np.: „Samotrzeciej”, „samoprzyznaniem”, „samorozwiązania”, „samorozwiązanie”, „zekowaniem”.
  - TaKIPI zmienia wielkość liter przy lematyzacji, np. lematyzuje „XVII-wieczny” jako „xvii-wieczny”.
  - SAM zmienia wielkość liter i usuwa myślniki przy lematyzacji, np. lematyzuje „XVII-wieczny” jako „xviiwieczny”.

Walidacja: formy poprawne spoza SGJP

	Liczba unikalnych form	Liczba form	Procent unikalnych form	Procent form
OK	14747	38816	88,9338%	91,9898%
OK CC	207	231	1,2483%	0,5474%
GOODPOS	151	177	0,9106%	0,4195%
GOODPOS CC	364	474	2,1952%	1,1233%
LEMMA	790	1383	4,7642%	3,2776%
LEMMA CC	181	935	1,0915%	2,2158%
FAIL	142	180	0,8564%	0,4266%
cały korpus	16582	42196	100,0000%	100,0000%

Oznaczenia:

- OK — przykład poprawnie przetworzony
- GOODPOS — zgodność lematu i części mowy
- LEMMA — zgodność lematu
- FAIL — brak zgodności
- CC — ignorowanie wielkości liter przy porównywaniu lematów

Walidacja: formy z dywizem i apostrofem

	Liczba unikalnych form	Liczba form	Procent unikalnych form	Procent form
OK	459	576	69,6510%	73,5632%
OK CC	3	3	0,4552%	0,3831%
GOODPOS	15	15	2,2762%	1,9157%
GOODPOS CC	2	2	0,3035%	0,2554%
LEMMA	23	23	3,4901%	2,9374%
FAIL	157	164	23,8240%	20,9451%
cały korpus	659	783	100,0000%	100,0000%

Oznaczenia:

- OK — przykład poprawnie przetworzony
- GOODPOS — zgodność lematu i części mowy
- LEMMA — zgodność lematu
- FAIL — brak zgodności
- CC — ignorowanie wielkości liter przy porównywaniu lematów

Walidacja: pt lematyzowane do sg przez SGJP

	Liczba unikalnych form	Liczba form	Procent unikalnych form	Procent form
OK	130	371	77,3810%	80,4772%
GOODPOS	13	13	7,7381%	2,8200%
GOODPOS CC	8	57	4,7619%	12,3644%
LEMMA CC	2	2	1,1905%	0,4338%
FAIL	15	18	8,9286%	3,9046%
cały korpus	168	461	100,0000%	100,0000%

Oznaczenia:

- OK — przykład poprawnie przetworzony
- GOODPOS — zgodność lematu i części mowy
- LEMMA — zgodność lematu
- FAIL — brak zgodności
- CC — ignorowanie wielkości liter przy porównywaniu lematów

Walidacja: tag inny niż proponowany przez SGJP

	Liczba unikalnych form	Liczba form	Procent unikalnych form	Procent form
OK	537	6261	46,6551%	56,8149%
OK CC	25	45	2,1720%	0,4083%
GOODPOS	61	157	5,2997%	1,4247%
GOODPOS CC	14	46	1,2163%	0,4174%
LEMMA	332	3512	28,8445%	31,8693%
LEMMA CC	105	708	9,1225%	6,4247%
FAIL	77	291	6,6898%	2,6407%
cały korpus	1151	11020	100,0000%	100,0000%

Oznaczenia:

- OK — przykład poprawnie przetworzony
- GOODPOS — zgodność lematu i części mowy
- LEMMA — zgodność lematu
- FAIL — brak zgodności
- CC — ignorowanie wielkości liter przy porównywaniu lematów

Walidacja: formy gwarowe bądź archaiczne

	Liczba unikalnych form	Liczba form	Procent unikalnych form	Procent form
OK	25	28	18,9394%	16,8675%
GOODPOS	5	5	3,7879%	3,0120%
GOODPOS CC	1	1	0,7576%	0,6024%
LEMMA	8	10	6,0606%	6,0241%
LEMMA CC	1	1	0,7576%	0,6024%
FAIL	92	121	69,6970%	72,8916%
cały korpus	132	166	100,0000%	100,0000%

Oznaczenia:

- OK — przykład poprawnie przetworzony
- GOODPOS — zgodność lematu i części mowy
- LEMMA — zgodność lematu
- FAIL — brak zgodności
- CC — ignorowanie wielkości liter przy porównywaniu lematów

Porównanie odgadywaczy: formy poprawne spoza SGJP

	Procent unikalnych form			Procent form		
	ENIAM	TaKIPI	SAM	ENIAM	TaKIPI	SAM
OK	88,93%	5,07%	0,00%	91,99%	48,12%	0,00%
OK CC	1,25%	13,10%	0,00%	0,55%	6,70%	0,00%
GOODPOS	0,91%	6,46%	8,45%	0,42%	2,96%	3,80%
GOODPOS_NONINFL	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS_VERB	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS CC	2,20%	27,66%	54,06%	1,12%	13,85%	26,39%
GOODPOS_NONINFL CC	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS_VERB CC	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
LEMMA	4,76%	6,69%	5,44%	3,28%	3,53%	5,93%
LEMMA CC	1,09%	7,18%	1,32%	2,22%	3,36%	0,56%
FAIL	0,86%	33,23%	30,69%	0,43%	21,23%	63,30%
CRASH	0,00%	0,00%	0,03%	0,00%	0,00%	0,01%

- OK — przykład poprawnie przetworzony
- GOODPOS — zgodność lematu i części mowy
- GOODPOS_NONINFL — zgodność lematu i tego, że część mowy jest nieodmienna
- GOODPOS_VERB — zgodność lematu i tego, że część mowy jest czasownikiem
- LEMMA — zgodność lematu
- FAIL — brak zgodności
- CRASH — runtime error
- CC — ignorowanie wielkości liter przy porównywaniu lematów

Porównanie odgadywaczy: formy z dywizem i apostrofem

	Procent unikalnych form			Procent form		
	ENIAM	TaKIPI	SAM	ENIAM	TaKIPI	SAM
OK	69,65%	19,42%	0,00%	73,56%	18,26%	0,00%
OK CC	0,46%	3,49%	0,00%	0,38%	3,70%	0,00%
GOODPOS	2,28%	21,24%	0,61%	1,92%	21,58%	0,51%
GOODPOS_NONINFL	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS_VERB	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS CC	0,30%	3,34%	7,28%	0,26%	2,94%	6,39%
GOODPOS_NONINFL CC	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS_VERB CC	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
LEMMA	3,49%	0,91%	0,00%	2,94%	0,77%	0,00%
LEMMA CC	0,00%	0,15%	0,00%	0,00%	0,13%	0,00%
FAIL	23,82%	46,43%	92,11%	20,95%	48,40%	93,10%
CRASH	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%

- OK — przykład poprawnie przetworzony
- GOODPOS — zgodność lematu i części mowy
- GOODPOS_NONINFL — zgodność lematu i tego, że część mowy jest nieodmienna
- GOODPOS_VERB — zgodność lematu i tego, że część mowy jest czasownikiem
- LEMMA — zgodność lematu
- FAIL — brak zgodności
- CRASH — runtime error
- CC — ignorowanie wielkości liter przy porównywaniu lematów

Porównanie odgadywaczy: pt lematyzowane do sg przez SGJP

	Procent unikalnych form			Procent form		
	ENIAM	TaKIPI	SAM	ENIAM	TaKIPI	SAM
OK	77,38%	1,19%	0,00%	80,48%	0,43%	0,00%
OK CC	0,00%	0,60%	0,00%	0,00%	0,22%	0,00%
GOODPOS	7,74%	0,60%	25,00%	2,82%	0,22%	13,23%
GOODPOS_NONINFL	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS_VERB	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS CC	4,76%	0,00%	25,00%	12,36%	0,00%	11,50%
GOODPOS_NONINFL CC	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS_VERB CC	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
LEMMA	0,00%	1,19%	0,60%	0,00%	0,43%	0,22%
LEMMA CC	1,19%	0,00%	0,00%	0,43%	0,00%	0,00%
FAIL	8,93%	96,43%	49,40%	3,90%	98,70%	75,05%
CRASH	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%

- OK — przykład poprawnie przetworzony
- GOODPOS — zgodność lematu i części mowy
- GOODPOS_NONINFL — zgodność lematu i tego, że część mowy jest nieodmienna
- GOODPOS_VERB — zgodność lematu i tego, że część mowy jest czasownikiem
- LEMMA — zgodność lematu
- FAIL — brak zgodności
- CRASH — runtime error
- CC — ignorowanie wielkości liter przy porównywaniu lematów

Porównanie odgadywaczy: tag inny niż proponowany przez SGJP

	Procent unikalnych form			Procent form		
	ENIAM	TaKIPI	SAM	ENIAM	TaKIPI	SAM
OK	46,66%	9,73%	0,00%	56,81%	27,30%	0,00%
OK CC	2,17%	0,00%	0,00%	0,41%	0,00%	0,00%
GOODPOS	5,30%	13,12%	13,03%	1,42%	8,08%	1,77%
GOODPOS_NONINFL	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS_VERB	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS CC	1,22%	0,09%	0,00%	0,42%	0,01%	0,00%
GOODPOS_NONINFL CC	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS_VERB CC	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
LEMMA	28,84%	35,27%	41,79%	31,87%	54,49%	64,95%
LEMMA CC	9,12%	0,00%	0,87%	6,42%	0,00%	0,59%
FAIL	6,69%	41,70%	44,31%	2,64%	10,11%	32,70%
CRASH	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%

- OK — przykład poprawnie przetworzony
- GOODPOS — zgodność lematu i części mowy
- GOODPOS_NONINFL — zgodność lematu i tego, że część mowy jest nieodmienna
- GOODPOS_VERB — zgodność lematu i tego, że część mowy jest czasownikiem
- LEMMA — zgodność lematu
- FAIL — brak zgodności
- CRASH — runtime error
- CC — ignorowanie wielkości liter przy porównywaniu lematów

Porównanie odgadywaczy: formy gwarowe bądź archaiczne

	Procent unikalnych form			Procent form		
	ENIAM	TaKIPI	SAM	ENIAM	TaKIPI	SAM
OK	18,94%	2,27%	0,00%	16,87%	1,81%	0,00%
OK CC	0,00%	0,76%	0,00%	0,00%	0,60%	0,00%
GOODPOS	3,79%	4,55%	4,55%	3,01%	4,82%	4,82%
GOODPOS_NONINFL	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS_VERB	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS CC	0,76%	0,76%	0,00%	0,60%	1,20%	0,00%
GOODPOS_NONINFL CC	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
GOODPOS_VERB CC	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
LEMMA	6,06%	1,52%	5,30%	6,02%	1,20%	4,22%
LEMMA CC	0,76%	0,00%	0,76%	0,60%	0,00%	1,20%
FAIL	69,70%	90,15%	89,39%	72,89%	90,36%	89,76%
CRASH	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%

- OK — przykład poprawnie przetworzony
- GOODPOS — zgodność lematu i części mowy
- GOODPOS_NONINFL — zgodność lematu i tego, że część mowy jest nieodmienna
- GOODPOS_VERB — zgodność lematu i tego, że część mowy jest czasownikiem
- LEMMA — zgodność lematu
- FAIL — brak zgodności
- CRASH — runtime error
- CC — ignorowanie wielkości liter przy porównywaniu lematów

Wersja demonstracyjna

- Przedstawiony w artykule model został zaimplementowany i stanowi fragment kategorialnego parsera składniowo-semantycznego „ENIAM”.
- Internetowa wersja demonstracyjna guessera dostępna jest pod adresem:
  `http://eniam.nlp.ipipan.waw.pl/morphology.html`.
- Internetowa wersja demonstracyjna generatora form dostępna jest pod adresem:
  `http://eniam.nlp.ipipan.waw.pl/morphology2.html`.

Kod źródłowy i zasoby

- Kod źródłowy, dane modelu i otagowana lista frekwencyjna NKJP1M znajdują się w repozytorium:
  `http://git.nlp.ipipan.waw.pl/wojciech.jaworski/ENIAM`
- Odpowiednio w
  - katalogu `morphology`,
  - katalogu `morphology/data` i
  - pliku `resources/NKJP1M/NKJP1M-tagged-frequency-26.07.2017.tab`
- Definicja tagsetu listy frekwencyjnej znajduje się w pliku
  `resources/NKJP1M/NKJP-tagged-frequency-tagset.txt`