

Prace, zasoby i narzędzia analizy lingwistycznej Zespołu Inżynierii Lingwistycznej IPI PAN



Agnieszka Mykowiecka | Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Warszawa, 30 października 2017

Przetwarzanie języka naturalnego



Odyseja kosmiczna 2001

analiza morfologiczna

0	1	Jan	Ewangelia_świątego_Jana	brev:npun	
		Jan		subst:sg:nom:m1	imię
1	2	zapalił	zapalić	praet:sg:m1.m2.m3:perf	
2	3	świecę	świeca	subst:sg:acc:f	nazwa posp
		świecić		fin:sg:pr:imperf	
3	4	.	.	interp	

ujednoznacznianie morfologiczne

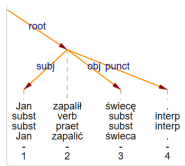
```

<orth>Jan</orth>
<lex disamb="1"><base>Jan</base><ctag>subst:sg:nom:m1</ctag></lex>
</tok>
<tok>
<orth>zapalił</orth>
<lex disamb="1"><base>zapalić</base><ctag>praet:sg:m1:perf</ctag></lex>
</tok>
<tok>
<orth>świecę</orth>
<lex disamb="1"><base>świeca</base><ctag>subst:sg:acc:f</ctag></lex>
</tok>

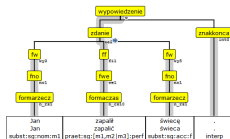
```

struktura syntaktyczna

drzewo zależnościowe

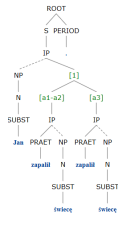


drzewo składnikowe

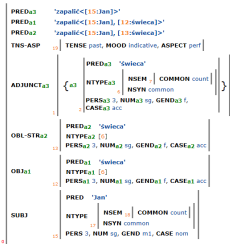


struktura LFG

C-structure



F-structure



Tematy badawcze

- ▶ korpusy języka polskiego — NKJP,
- ▶ analiza morfoskładniowa — SGJP, Polimorf, Morfeusz,
- ▶ tagowanie morfologiczne – Concraft
- ▶ analiza składniowa (powierzchniowa i głęboka) — Świgr, parser zależnościowy, gramatyka LFG,
- ▶ rozpoznawanie nazw własnych — Nerf,
- ▶ identyfikacja terminów — TermoPL,
- ▶ analiza sentymentu, opinii,
- ▶ analiza zależności referencyjnych,
- ▶ analiza zależności czasowych, identyfikacja opisu sytuacji w tekście
- ▶ semantyka dystrybucyjna
- ▶ ...

Projekty ZIL

- ▶ **CLARIN-PL, DARIAH-PL** — budowa infrastruktury informatycznej wspierającej badaczy-humanistów,
- ▶ **Chronofleks, KORBA** — analiza tekstów dawnych,
- ▶ **Swad, CoDeS** — metody semantyki dystrybucyjnej,
- ▶ **Cothec** — Ujednolicona teoria koreferencji w języku polskim i jej korpusowa weryfikacja,
- ▶ **TextLink** — Structuring Discourse in Multilingual Europe.

<http://zil.ipipan.waw.pl/>

Korpusy językowe

- ▶ zbiory tekstów z różnym rodzajem oznaczeń dodatkowych
- ▶ potencjalne prace:
 - ▶ kolekcjonowanie tekstów,
 - ▶ rozbudowywanie anotacji,
 - ▶ ewaluacja, korekta,
 - ▶ wykorzystanie w badaniach.

Korpusy językowe

- ▶ NKJP, <http://nkjp.pl/> (aktualnie nierozwijany)

 **NARODOWY KORPUS
JĘZYKA POLSKIEGO**

KONSORCJUM


O PROJEKCIE NKJP
ZESPÓŁ
PUBLIKACJE
SŁOWA DNIA
TEKSTY KORPUSU
ZASTOSOWANIA
PODZIĘKOWANIA
NARZĘDZIA I ZASOBY
KONTAKT

O projekcie NKJP

Korpus językowy to zbiór tekstów, w którym szukamy typowych użyć słów i konstrukcji oraz innych informacji o ich znaczeniu i funkcji. Bez dostępu do korpusu nie da się dziś prowadzić badań językoznawczych, pisać słowników ani podręczników języków obcych, tworzyć wyszukiwarek uwzględniających polską odmianę, tłumaczy komputerowych ani innych programów zaawansowanej technologii językowej. Korpus jest niezbędny do pracy językoznawcom, ale korzystają zeń często także informatycy, historycy, bibliotekarze, badacze literatury i kultury oraz specjaliści z wielu innych dziedzin humanistycznych i informatycznych.

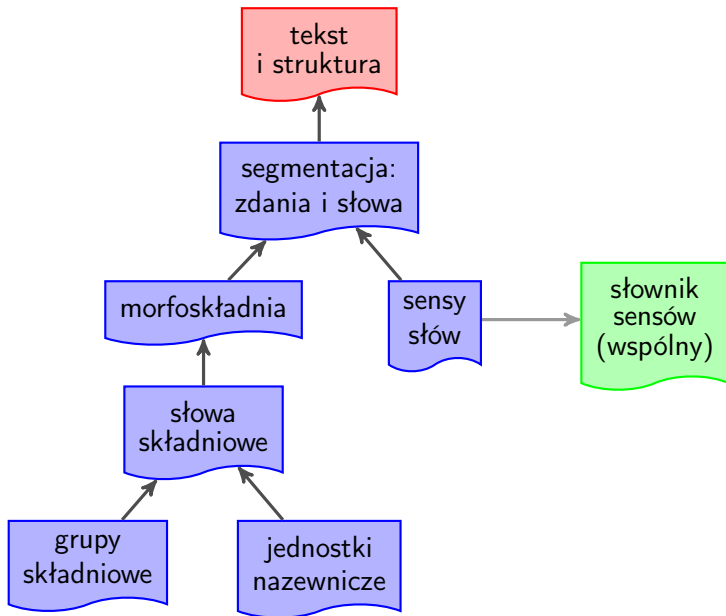
Swoje korpusy narodowe mają już Brytyjczycy, Niemcy, Czesi i Rosjanie. Także Polakom potrzebny jest wielki, zrównoważony gatunkowo i tematycznie, korpus językowy – internetowy skarbiec polszczyzny.

**WYSZUKIWARKA KORPUSOWA
IPI PAN**

**WYSZUKIWARKA KORPUSOWA
PELCRA**

- ▶ korpus parlamentarny ...

NKJP, warstwowy opis polszczyzny



Korpus parlamentarny

- ▶ duży zbiór stenogramów, interpelacji i zapytań, obejmujący okres od 1991 r. (<http://clip.ipipan.waw.pl/PPC>)
- ▶ anotowany lingwistycznie i dostępny do przeszukiwania (<http://sejm.nlp.ipipan.waw.pl/>)
- ▶ aktualnie rozszerzany o teksty stenogramów od 1918 r. (okres międzywojenny + PRL) oraz zapisy z posiedzeń komisji (OCR + ręczna korekta).

Wyszukiwanie w korpusach

NKJP

1.	przez ten czas nie dała	<u>odczuć</u> [odczucie:subst:pl:gen:n]	gościowi niechęci. "Dla
2.	jest w stanie lepiej ją	<u>odczuć</u> [odczucie:subst:pl:gen:n]	, zrozumieć, przekazać i
3.	i zanim duch cokolwiek znowu	<u>odczuć</u> [odczucie:subst:pl:gen:n]	zdoła, zmysły okropnie mnie
4.	laptopami, płaskimi komórkami płaskich	<u>odczuć</u> [odczucie:subst:pl:gen:n]	ułożonych na płask w spłaszczonych
5.	, że jest w stanie	<u>odczuć</u> [odczucie:subst:pl:gen:n]	kamienną niecierpliwość, choć nawet
6.	bardzo blisko mnie, żeby	<u>odczuć</u> [odczucie:subst:pl:gen:n]	ten nieustannie przepływający przeze mnie
7.	. Właśnie pozwoliłam sobie	<u>odczuć</u> [odczucie:subst:pl:gen:n]	głód rodziny, jej ograniczeń
8.	by choć rąbkami świadomości nie	<u>odczuć</u> [odczucie:subst:pl:gen:n]	przez mgnienie swojej przewagi nad
9.	bezduszny, obojętny człowiek mógł	<u>odczuć</u> [odczucie:subst:pl:gen:n]	jego wściekłość. Jego nienawiść
10.	nie chciałem negocjować pani	<u>odczuć</u> [odczucie:subst:pl:gen:n]	i życiowych doświadczeń. Zastanawiał

Sejm

1.	przekonany, iż będę wyrazicielem	<u>odczuć</u> [odczucie:subst:pl:gen:n]	wielu pań i panów posłów
2.	. W Europie daje się	<u>odczuć</u> [odczucie:subst:pl:gen:n]	wrogość i uprzedzenia do chrześcijan
3.	A pacjenci szybko powinni to	<u>odczuć</u> [odczucie:subst:pl:gen:n]	, tak jak w już
4.	NATO. Nie sposób nie	<u>odczuć</u> [odczucie:subst:pl:gen:n]	pokory, kiedy porównujemy to
5.	w Europie. Mogę to	<u>odczuć</u> [odczucie:subst:pl:gen:n]	w wielu miejscach, w
6.	rzeczywiście w Polsce można było	<u>odczuć</u> [odczuć:inf:perf]	wyraźną poprawę w stosunkach pracy
7.	, to polskie przedsiębiorstwa mogą	<u>odczuć</u> [odczuć:inf:perf]	znaczne pogorszenie się sytuacji,
8.	kryzysu dają się coraz bardziej	<u>odczuć</u> [odczuć:inf:perf]	również w naszej gospodarce.
9.	To jest kwestia bardzo subiektywnych	<u>odczuć</u> [odczuć:inf:perf]	. Skoro jednak i telewizja
10.	tej sali, można było	<u>odczuć</u> [odczucie:subst:pl:gen:n]	, jak potężną siłą jest

Korpus parlamentarny – proponowane tematy prac

- ▶ konwersja tekstów po korekcie i bieżących posiedzeń komisji do formatu TEI P5,
- ▶ powiązanie tekstów z ontologią polityczną (<http://zil.ipipan.waw.pl/PolishPoliticalOntology>),
- ▶ stworzenie "żywego" korpusu przetworzonego najlepszymi narzędziami do analizy lingwistycznej,
- ▶ opracowanie webserwisów udostępniających dane korpusu,
- ▶ implementacja ngram viewera na bazie wszystkich danych parlamentarnych,
- ▶ udział w stworzeniu serwisu udostępniającego parlamentarne dane lingwistyczne,
- ▶ udział w pracach nad korpusem aktów prawnych.

KORPUS: W1

STAN: NOWY

Nazwa tekstu		Autor	Liczba segmentów	Udział	Stan	Operacja
zakazane-miasto.txt		wikipedia	843	100,0%	Gotowy	Pobierz tekst Edytuj metadane Usuń



Korpusomat – tematy prac

- ▶ wizualizacja informacji o korpusie tekstowym w Korpusomacie,
- ▶ słowa tematyczne (częste słowa po odrzuceniu stopwords),
- ▶ słowa kluczowe (słowa charakterystyczne dla danego korpusu),
- ▶ rozkład częstości słów w korpusie.
- ▶ terminy w korpusie (integracja z TermoPL).
- ▶ "word sketches"
(https://www.researchgate.net/profile/Milos_Jakubicek/publication/)

strategy (noun)
English Corpus for SkELL freq = 117,755 (79.07 per million)

<u>modifiers of strategy</u>	<u>95,790</u>	<u>2.00</u>	<u>verbs with strategy as object</u>	<u>41,483</u>	<u>2.30</u>	<u>words and/or strategy</u>	<u>22,751</u>	<u>1.10</u>	<u>verbs with strategy as subject</u>
marketing	<u>2,143</u>	8.64	implement	<u>1,271</u>	8.37	tactic	<u>910</u>	8.70	backfire
effective	<u>1,113</u>	7.14	cope	<u>468</u>	8.31	objective	<u>196</u>	6.04	aim
overall	<u>791</u>	7.04	devise	<u>431</u>	7.85	marketing	<u>205</u>	6.02	involve
long-term	<u>610</u>	7.00	develop	<u>2,848</u>	7.61	planning	<u>191</u>	5.81	focus
investment	<u>789</u>	6.75	adopt	<u>995</u>	7.51	policy	<u>650</u>	5.56	depend
exit	<u>400</u>	6.74	formulate	<u>271</u>	7.28	technique	<u>314</u>	5.54	fail

Analiza morfologiczna

0	1	<i>Co</i>	co	subst:sg:nom.acc:n2
1	2	<i>ś</i>	BYĆ	aglt:sg:sec:imperf:nwok
0	2	<i>Coś</i>	coś	subst:sg:nom.acc:n2
2	3	<i>zrobił</i>	ZROBIĆ	praet:sg:m1.m2.m3:perf
3	4	<i>?</i>	<i>?</i>	interp

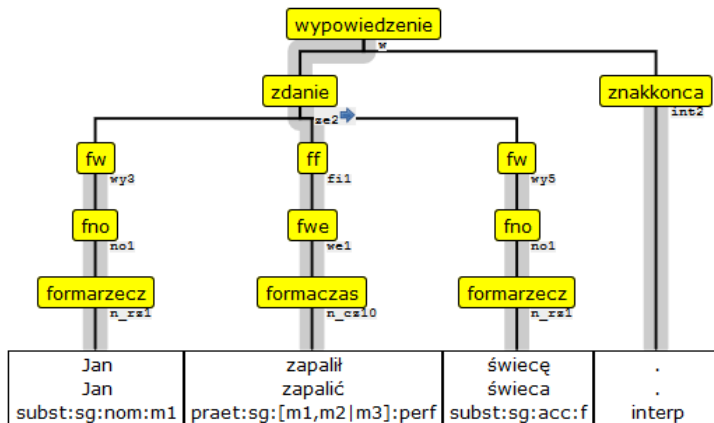
Opracowanie modułu Pythona dla analizatora fleksyjnego Morfeusz 2 (<http://sgjp.pl/morfeusz/>)

- ▶ dający się użyć wieloplatformowo (z wykorzystaniem biblioteki ctypes)
- ▶ pozwalającego na dostęp do wszystkich możliwości tej wersji Morfeusza.

Wymagania: dobra znajomość Pythona

Marcin Woliński

Parsowanie składniowe, drzewo fraz, Świga



Parsowanie składniowe, Świga

Napisać moduł wizualizujący drzewa składniowe generowane przez parser Świga.

- ▶ Wzorcowy moduł (ale działający dobrze tylko w Firefoxie) jest, zob. <http://swigra.nlp.ipipan.waw.pl/>
- ▶ Zadanie polegałoby na tym, żeby skonstruować wizualizację dającą wszystkie możliwości dotychczasowej, ale działającą w dowolnej przeglądarce (no, powiedzmy Firefox, IE, Safari, Opera?).
- ▶ Byłoby też dobrze pozwolić na wizualizację drzew składniowych z nieciągłościami i pozwolić, żeby zestawy atrybutów węzłów mogły być zagnieżdżonymi strukturami atrybutów, a nie tylko listą.

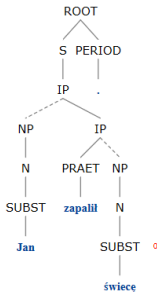
Wymagania: znajomość JavaScriptu

Marcin Woliński

Parsowanie składniowe, struktury LFG



C-structure



F-structure

PRED	'zapalić<[10:Jan], [7:świeca]>'
TNS-ASP	TENSE past, MOOD indicative, ASPECT perf
OBJ	PRED 'świeca' NTYPE NSEM 5 COMMON count 4 NSYN common 7 PERS 3, NUM sg, GEND f, CASE acc
SUBJ	PRED 'Jan' NTYPE NSEM 13 COMMON count 12 NSYN common 10 PERS 3, NUM sg, GEND m1, CASE nom

Parsowanie składniowe, struktury LFG

Wytrenowanie parsera zależnościowego na podstawie istniejącego banku struktur LFG.

- ▶ konwersja banku struktur LFG na zależnościowy bank struktur, z uwzględnieniem różnych podejść do tego problemu (m.in. konwersja na UD (Universal Dependencies) i nie-UD).
- ▶ trenowanie parsera zależnościowego na podstawie otrzymanego banku struktur zależnościowych

Wymagania: biegłość programistyczna, pewna wiedza lingwistyczna

Adam Przepiórkowski, Agnieszka Patejuk

Identyfikacja informacji w tekście – relacje czasowe

Žarko Čabarkapa

Žarko Čabarkapa, (ur. ⁿ¹ 21 maja 1981 roku w Belgradzie), serbski koszykarz, występujący w barwach KK Budućnost Podgorica, reprezentant Serbii. Ma 211 cm wzrostu, waży 106 kg.

Kariera

Przed ⁿ² przybyciem do NBA występował ^{n2 n4 n3 n5 n6} w serbskich drużynach Bleopetrol (^{n3 n1} 1998 - ⁿ¹ 01) i Budućnost (ⁿ³ 2001 - ^{n3 n3} 03). W ⁿ² 2003 roku został wybrany ^{n7 n2} przez Phoenix Suns z numerem 17. w draftcie. W swoim pierwszym sezonie w NBA w barwach Suns ⁿ⁸ wystąpił w 49 meczach ⁿ⁹ zdobywając średnio 4.1 punktu na mecz. W ⁿ² drugim roku gry w NBA (ⁿ¹⁰ 2004 - ⁿ¹¹ 05) ⁿ¹¹ rozegrał ^{n8 n9 n10} w barwach Suns zaledwie 3 mecze po czym ⁿ¹² przeniósł ⁿ¹¹ się do Golden State Warriors. Tam ⁿ¹³ wystąpił ⁿ¹² aż w 77 spotkaniach ⁿ¹³ zdobywając 6.1 punktu na mecz. W ⁿ¹⁴ sezonie 2005 - ⁿ¹⁵ 06 w barwach Warriors ⁿ¹⁵ rozegrał ^{n13 n14} 61 spotkań ⁿ¹⁵ zrucając 3.3 punktu na mecz. ⁿ¹⁶ 16 stycznia 2009 ⁿ¹⁶ podpisał kontrakt z KK Budućnost Podgorica.

Annotations			
Stage: new final discarded			
Id	Text	From-to	Type
283672	21 maja 1981 roku	33-46	t3_date
283673	1998	228-231	t3_date
283674	01	233-234	t3_date
283675	2001	247-250	t3_date
283676	03	252-253	t3_date
283677	2003 roku	257-264	t3_date
283691	2004	422-425	t3_date
283692	05	427-428	t3_date

Relations		
LINK_AF-1	rozegrał	wystąpił
TLINK_AFT	przeniósł	rozegrał
TLINK_AFT	wystąpił	przeniósł
TLINK_AFT	rozegrał	wystąpił
TLINK_BEf	występował	przybyciem
TLINK_BEG_BY	występował	1998
TLINK_DUR	występował	1998-01
TLINK_DUR	występował	2001-03
TLINK_DUR	rozegrał	drugim roku
TLINK_DUR	rozegrał	2004-05
TLINK_DUR	rozegrał	sezonie 2005-06

Identyfikacja informacji w tekście – relacje czasowe

Wytrenowanie modułu rozpoznającego wyrażenia temporalne i informacje o relacjach temporalnych na podstawie oznaczonego zbioru danych.

Agnieszka Mykowiecka, Małgorzata Marciniak

Semantyka dystrybucyjna

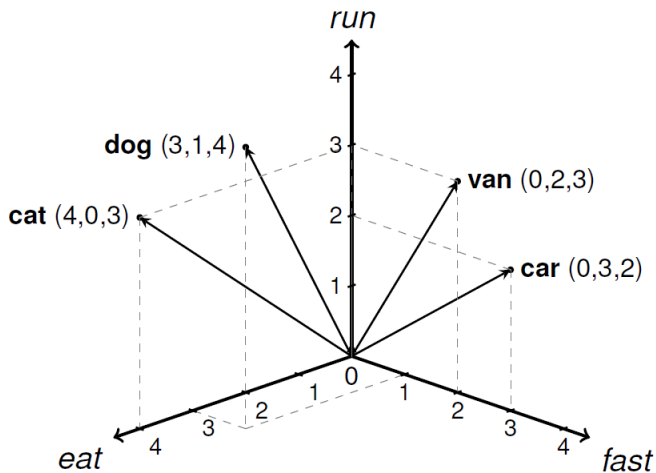
Leksemy o podobnych cechach dystrybucyjnych mają podobne znaczenie

Wektory współwystępowania

	<i>dog</i>	<i>drive</i>	<i>eat</i>	<i>fast</i>	<i>play</i>	<i>...</i>	<i>the</i>	<i>wheel</i>
car	0	3	0	2	0	$\vdots$	2	1
cat	1	0	3	0	1	$\vdots$	2	0
dog	0	0	3	0	2	$\vdots$	2	0
van	0	3	0	1	0	$\vdots$	3	1

co-occurrence matrix

Wektory współwystępowania



Szukanie słów podobnych

dom plWordnet

podobieństwo jednostka leksykalna

0.193	domek	✓
0.186	budynek	✓
0.173	mieszkanie	✓
0.152	kamienica	✓
0.141	domostwo	✓
0.141	rodzina	
0.141	rezydencja	
0.139	chata	✓
0.129	chałupa	✓
0.126	posiadłość	
0.125	pokój	
0.123	pałac	✓
0.120	dwór	✓
0.118	budowla	
0.115	hotel	
0.115	stodoła	
0.110	gmach	✓
0.109	garaż	
0.109	posesja	
0.108	sklep	

NKJP+Wikipedia

domek	0.73	✓
chałupa	0.64	✓
barak	0.62	
chata	0.61	✓
mieszkanie	0.60	✓
willa	0.60	✓
dworek	0.58	✓
altanka	0.57	
gospoda	0.56	
sypialnia	0.56	
kamienica	0.55	✓
światlica	0.55	
domostwo	0.55	✓
Domek	0.55	
piwnica	0.55	
plebania	0.55	
schronisko	0.55	
pałacyk	0.54	
szatas	0.54	
chatka	0.54	✓
pensjonat	0.54	
suterena	0.53	
baraczek	0.53	
pałac	0.53	✓
altana	0.53	

Wikipedia

pałacyk	0.71
domek	0.70
dworek	0.68
pałac	0.67
pawilon	0.66
szpital	0.65
lokal	0.65
pensjonat	0.64
budynek	0.64
apartament	0.62
sierociniec	0.62
dwór	0.60
grobowiec	0.60
pasaż	0.59
zajazd	0.58
ogród	0.58
zameczek	0.58
kościół	0.57
klasztor	0.56
lamus	0.56
sklep	0.56
przysiółek	0.55
konwikt	0.55
żłobek	0.55
ośrodek	0.55

Semantyka dystrybucyjna – tematy

- ▶ identyfikacja słów powiązanych wybraną relacją, np. synonimia, hyperonimia, co-hyponimia
- ▶ rozpoznawanie znaczenia słowa w kontekście (ew. prace anotacyjne)
- ▶ wytrenowanie dystrybucyjnych modeli kompozycyjnych, np. macierzy dla przymiotników
- ▶ inne tematy do uzgodnienia

Modele, demo: <http://dsmodels.nlp.ipipan.waw.pl/analogy.html>

projekty: zil.ipipan.waw.pl/CoDeS, zil.ipipan.waw.pl/Scwad

Agnieszka Mykowiecka, Małgorzata Marciniak
Alina Wróblewska

Przypisanie znaczenia słowu w kontekście

WSD - word sense disambiguation

- problem 1: trudność, trudna sytuacja, z którą człowiek musi się zmierzyć
- problem 2: zagadnienie, zadanie do rozwiązania, kwestia do przemyślenia, podjęcia decyzji

Raport podkreśla, że w wielu krajach poważnym problemem/1 jest wysokie bezrobocie.

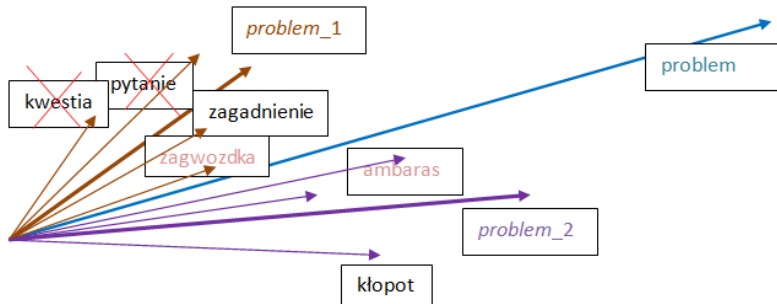
Pozostaje wreszcie problem/2 walorów dydaktycznych książki.

Przypisanie znaczenia słowu w kontekście

WSD - word sense disambiguation

problem_1: kwestia, pytanie, zagwozdzka, zagadnienie

problem_2: bieda, kłopot, ambaras



Semantyka dystrybucyjna – demo

Modele

Podobieństwo wyrazów

Analogie

Arytmetyka wyrazów

Analogie

Wybierz maksymalnie najlepszych rezultatów.

Użyj modelu opartego na: ☐ formach ☒ lematach.

Słowo 1:	Słowo 2:	Słowo 3:	Słowo 4:
mężczyzna	król	kobieta	królowa

#	Słowo	Stopień podobieństwa
1	królowa	0.699202775955
2	monarcha	0.696870446205
3	władca	0.692474961281
4	cesarz	0.664051294327
5	książę	0.613953709602
6	monarchini	0.581223011017
7	królewicz	0.573877990246
8	cesarzowa	0.572256147861
9	władczyni	0.572076141834
10	tron	0.563659071922