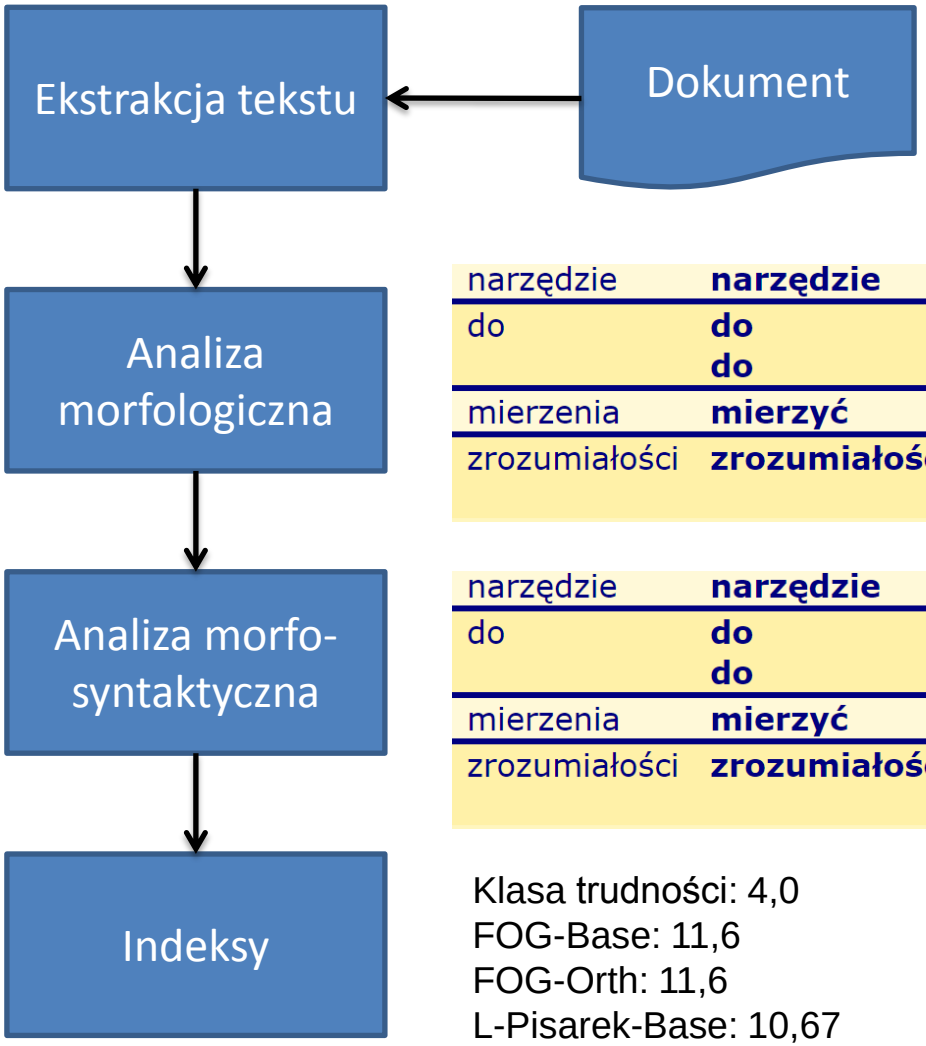
SWPS

Co wylicza Jasnopis?

Analiza języka polskiego



 narzędzie do mierzenia zrozumiałości

narzędzie	**narzędzie**	subst:sg:nom.acc.voc:n2
do	**do**	prep:gen
	do	subst:sg.pl:nom.gen.dat.acc.inst.loc.voc:n2
mierzenia	**mierzyć**	ger:sg:gen:n2:imperf:aff
zrozumiałości	**zrozumiałość**	subst:sg:gen.dat.loc.voc:f
		subst:pl:nom.gen.acc.voc:f

narzędzie	**narzędzie**	subst:sg:nom.acc.voc:n2
do	**do**	prep:gen
	do	~~subst:sg.pl:nom.gen.dat.acc.inst.loc.voc:n2~~
mierzenia	**mierzyć**	ger:sg:gen:n2:imperf:aff
zrozumiałości	**zrozumiałość**	subst:sg:gen.dat.loc.voc:f
		~~subst:pl:nom.gen.acc.voc:f~~

Klasa trudności: 4,0
FOG-Base: 11,6
FOG-Orth: 11,6
L-Pisarek-Base: 10,67
L-Pisarek-Orth: 10,67

Indeksy w Jasnopisie

- Indeks mglistości FOG
- Indeks Pisarka
- Automatyczne testy Taylora
- Grafy podobieństwa
- Dodatkowe statystyki
- Klasa trudności

Indeks FOG

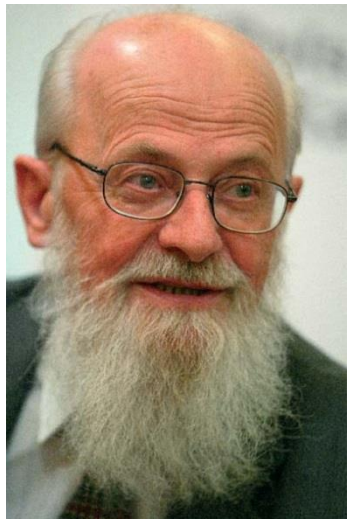
$$FOG = 0.4 \times \left(\frac{liczba\ wyrazów}{liczba\ zdań} + 100 \frac{liczba\ wyrazów\ trudnych}{liczba\ wyrazów} \right)$$

- Wyraz trudny: 4, lub więcej sylaby
- Granice zdań wyznaczone przez WCRFT
- Warianty operujące na
  - formach podstawowych wyrazów
  - formach ortograficznych
- Wygładzony indeks FOG wykorzystujący listy wyrazów łatwych
  - Lista Imiołczyka
  - Lista 5 tyś najczęstszych wyrazów

Interpretacja indeksu FOG

Wartość FOG	Interpretacja
1-6	język bardzo prosty, zrozumiały już dla uczniów szkoły podstawowej
7-9	język prosty, zrozumiały już dla uczniów gimnazjum
10-12	język dość prosty, zrozumiały już dla uczniów liceum
13-15	język dość trudny, zrozumiały dla studentów studiów licencjackich
16-17	język trudny, zrozumiały dla studentów studiów magisterskich
18 i więcej	język bardzo trudny, zrozumiały dla magistrów i osób z wyższym wykształceniem

Indeks Pisarka

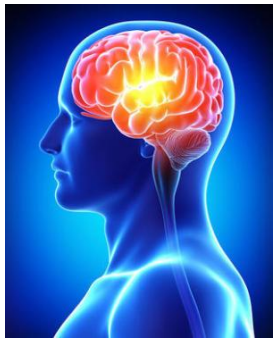


- Podobnie jak indeks FOG wykorzystuje
  - średnią długość zdania (ŚDZ)
  - procent wyrazów trudnych (PWT)
- Wersje: liniowa i nieliniowa
- Warianty operujące na
  - formach podstawowych wyrazów
  - formach ortograficznych
- Wygładzony indeks FOG wykorzystujący listy wyrazów łatwych
- Lista Imiołczyka
- Lista 5 tyś najczęstszych wyrazów

$$P_{NL} = \frac{1}{2}\sqrt{\acute{S}DZ^2 + PWT^2}$$

$$P_L = \frac{1}{3} \times \acute{S}DZ \times \frac{1}{3} \times PWT$$

Automatyczny test Taylora

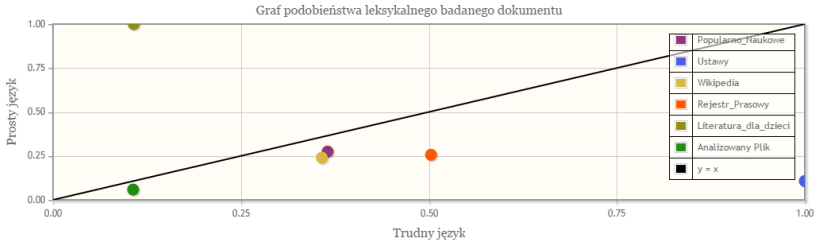


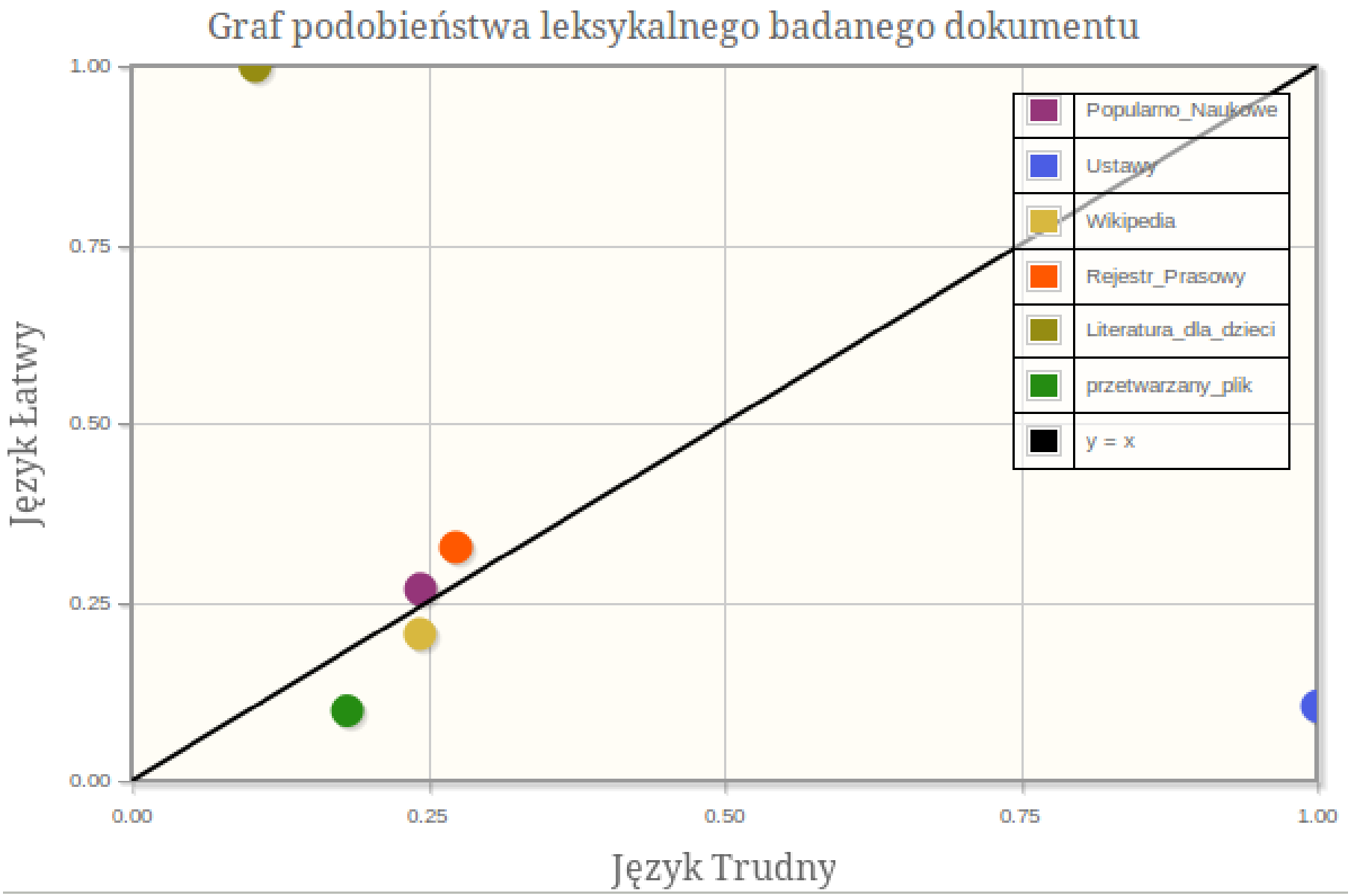
- Metoda Taylora – klasyczna metoda mierzenia czytelności poprzez uzupełnianie luk w tekście przez użytkowników języka
- Wytrenowanie modeli językowych na tekstach referencyjnych
- Uzupełnianie luk w tekście z wykorzystaniem modeli językowych
- Warianty:
  - Uzupełnianie co n-tego słowa poprzez model
  - Mierzenie odwrotności entropii (perplexity)

$$p(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_{w_i} c(w_{i-1}w_i)}$$

Grafy podobieństwa

- Podobieństwo pomiędzy korpusami referencyjnymi a tekstem użytkownika
- Worek słów
- Kosinus kąta pomiędzy wektorami jako miara podobieństwa
- Dwa modele porównywania tekstów
    - tf.idf
    - model binarny
- Porównanie na poziomie leksyki




Graf podobieństwa leksykalnego badanego dokumentu
Język Łatwy
Język Trudny
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Popularno_Naukowe
Ustawy
Wikipedia
Rejestr_Prasowy
Literatura_dla_dzieci
przetwarzany_plik
y = x

Weryfikacja

- Korpusy:
  - Literatura dziecięca (bajki)
  - Wikipedia
  - Artykuły prasowe (Rzeczpospolita)
  - Ustawy
  - Teksty popularno-naukowe (Wiedza i życie)
- Ok. 40 tyś słów/korpus dla podobieństwa
- Ok. 186 tyś słów/korpus dla automatycznego testu Taylora
- Walidacja krzyżowa

Weryfikacja

	Binarny	tf.idf
Literatura dla dzieci	100%	100%
Wikipedia	85,37%	85,37%
Ustawy	100%	100%
Artykuły prasowe	71,74%	73,91%
Popularno-naukowe	100%	100%

	Co 5 wyraz	Perplexity
Literatura dla dzieci	93,79%	97,18%
Wikipedia	80,56%	67,11%
Ustawy	86,29%	100%
Artykuły prasowe	71,66%	66,11%
Popularno-naukowe	73,77%	68,31%

Dziękuję za uwagę