

Indukcja reguł gramatyki języka polskiego

dr inż. Marcin Gołębski
m.golebski@elka.pw.edu.pl

Instytut Informatyki
Politechnika Warszawska

25 lutego 2008



Plan prezentacji

- 1 Aktualny stan wiedzy
 - Twierdzenie Golda
 - Pojęcia podstawowe
 - Prace Y. Sakakibary
 - Przetwarzanie wstępne tekstu
- 2 Nowe metody
 - Wyniki badań D. Magermana
 - Wyniki badań M. Gołębskiego
- 3 Rezultaty eksperymentów
- 4 Podsumowanie
 - Wyniki przeprowadzonych prac
 - Zastosowanie
 - Dalsze kierunki badań



POLITECHNIKA
WARSZAWSKA



Plan prezentacji

- 1 Aktualny stan wiedzy
 - Twierdzenie Golda
 - Pojęcia podstawowe
 - Prace Y. Sakakibary
 - Przetwarzanie wstępne tekstu
- 2 Nowe metody
 - Wyniki badań D. Magermana
 - Wyniki badań M. Gołębskiego
- 3 Rezultaty eksperymentów
- 4 Podsumowanie
 - Wyniki przeprowadzonych prac
 - Zastosowanie
 - Dalsze kierunki badań

Twierdzenie Golda (rok 1967)



E. M. Gold.

Language identification in the limit.

Information and Control, 10(5):447–474, 1967.

Twierdzenie

Klasa gramatyk zawierająca wszystkie języki skończone i przynajmniej jeden język nieskończony nie jest identyfikowalna wyłącznie z pozytywnych przykładów.



Plan prezentacji

- 1 Aktualny stan wiedzy
 - Twierdzenie Golda
 - Pojęcia podstawowe
 - Prace Y. Sakakibary
 - Przetwarzanie wstępne tekstu
- 2 Nowe metody
 - Wyniki badań D. Magermana
 - Wyniki badań M. Gołębskiego
- 3 Rezultaty eksperymentów
- 4 Podsumowanie
 - Wyniki przeprowadzonych prac
 - Zastosowanie
 - Dalsze kierunki badań

Pojęcia podstawowe

Gramatyka transformacyjna

$$(\Sigma, S, V, P)$$

gdzie:

Σ – alfabet,

S – symbol początkowy,

V – skończony zbiór symboli nieterminalnych takich, że

$$V \cap \Sigma = \emptyset,$$

$P \in (V \cup \Sigma)^* \times (V \cup \Sigma)^*$ – zbiór reguł produkcji.

Plan prezentacji

- 1 Aktualny stan wiedzy
 - Twierdzenie Golda
 - Pojęcia podstawowe
 - Prace Y. Sakakibary
 - Przetwarzanie wstępne tekstu
- 2 Nowe metody
 - Wyniki badań D. Magermana
 - Wyniki badań M. Gołębskiego
- 3 Rezultaty eksperymentów
- 4 Podsumowanie
 - Wyniki przeprowadzonych prac
 - Zastosowanie
 - Dalsze kierunki badań

Prace Y. Sakakibary (rok 1997)

Definicja

Gramatyka **reversible context-free** jest gramatyką bezkontekstową

$$G = (V, \Sigma, P, S)$$

taką, że:

- ❶ $A \rightarrow \alpha$ i $B \rightarrow \alpha$ znajdują się w zbiorze produkcji P implikuje, że $A = B$,
- ❷ $A \rightarrow \alpha B \beta$ i $A \rightarrow \alpha C \beta$ znajdują się w zbiorze produkcji P implikuje, że $B = C$,

gdzie A, B i C są symbolami nieterminalnymi, a $\alpha, \beta \in (N \cup \Sigma)^*$.



Prace Y. Sakakibary (rok 1997)

Twierdzenie

Klasa gramatyk **reversible context-free** jest identyfikowalna w granicy z użyciem pozytywnej reprezentacji ustrukturalizowanych łańcuchów, takich, które zostałyby wygenerowane przez gramatykę reversible context-free nieznanego języka bezkontekstowego.



Przykład ustrukturalizowanej reprezentacji zdania

Duży pies goni małą dziewczynkę

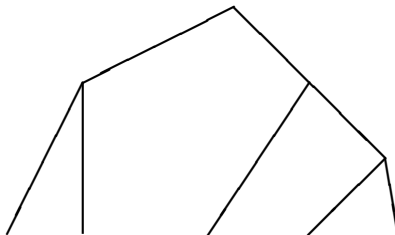


Przykład ustrukturalizowanej reprezentacji zdania

< <Duży pies> <goni <małą dziewczynkę> > >

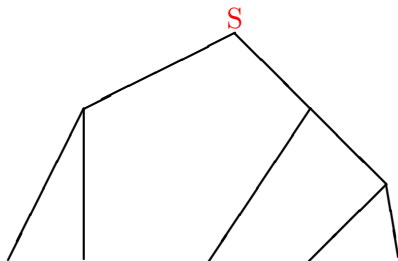


Przykład ustrukturalizowanej reprezentacji zdania



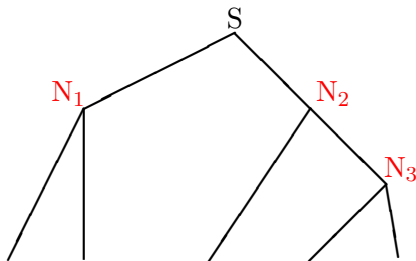
< <Duży pies> <goni <małą dziewczynkę> > >

Przykład ustrukturalizowanej reprezentacji zdania



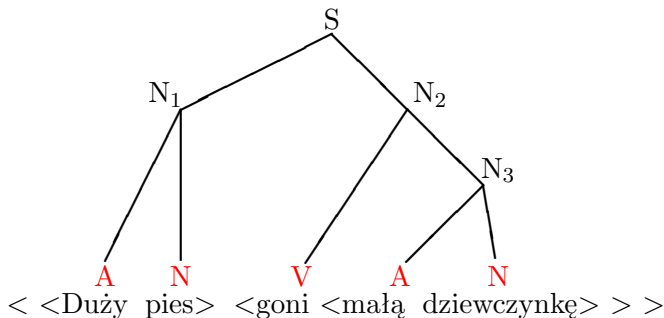
< <Duży pies> <goni <małą dziewczynkę> > >

Przykład ustrukturalizowanej reprezentacji zdania



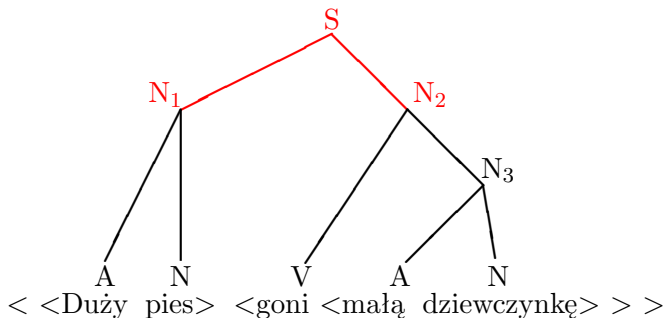
< <Duży pies> <goni <małą dziewczynkę> > >

Przykład ustrukturalizowanej reprezentacji zdania

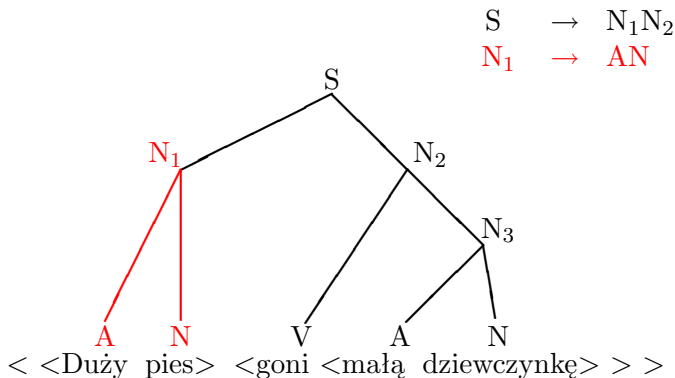


Przykład ustrukturalizowanej reprezentacji zdania

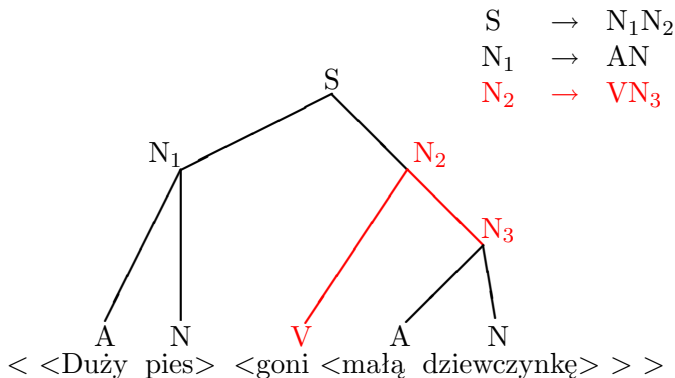
$$S \rightarrow N_1 N_2$$



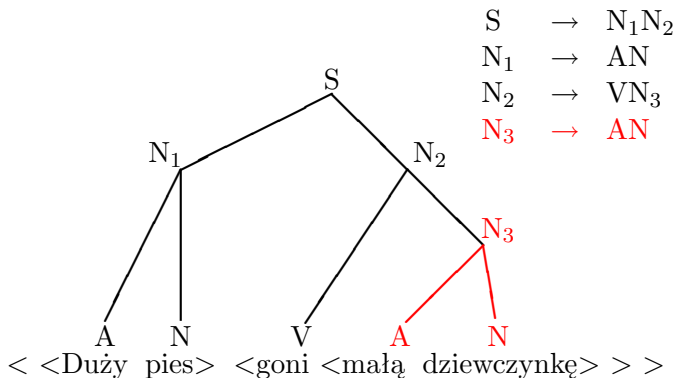
Przykład ustrukturalizowanej reprezentacji zdania



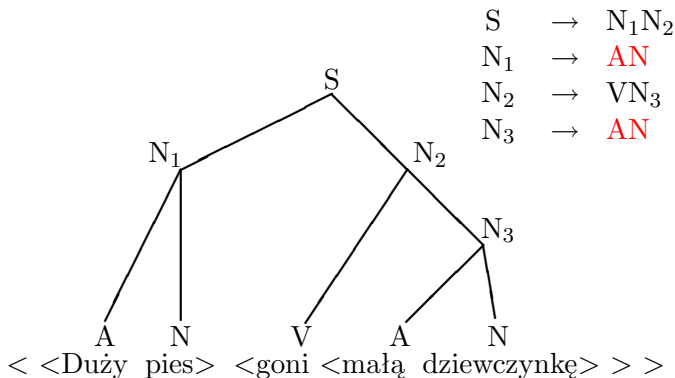
Przykład ustrukturalizowanej reprezentacji zdania



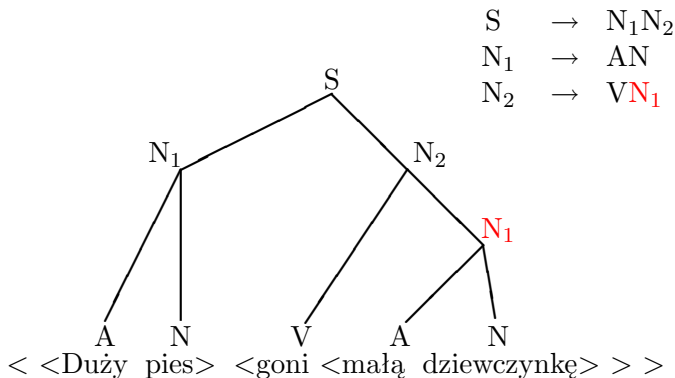
Przykład ustrukturalizowanej reprezentacji zdania



Przykład ustrukturalizowanej reprezentacji zdania



Przykład ustrukturalizowanej reprezentacji zdania



Plan prezentacji

- 1 Aktualny stan wiedzy
 - Twierdzenie Golda
 - Pojęcia podstawowe
 - Prace Y. Sakakibary
 - Przetwarzanie wstępne tekstu
- 2 Nowe metody
 - Wyniki badań D. Magermana
 - Wyniki badań M. Gołębskiego
- 3 Rezultaty eksperymentów
- 4 Podsumowanie
 - Wyniki przeprowadzonych prac
 - Zastosowanie
 - Dalsze kierunki badań

Przetwarzanie wstępne tekstu

- tokenizacja,
- wydzielanie zdań,
- analiza morfologiczna,
- ujenoznacznianie

Duży	pies	goni	małą	dziewczynkę
acc	m2	fin	acc	acc
adj	nom	imperf	adj	f
m3	sg	sg	pos	sg
pos	subst	ter	sg	subst
sg				

Informacja morfologiczna

Liczba etykiet

- w systemie CLAWS-5 — 76 etykiet
- w Korpusie IPI PAN — 2953 (778) etykiet

Metoda odwzorowania

- umożliwiająca redukcję liczby etykiet
- możliwy dobór ewolucyjny i manualny
- intuicyjna
- umożliwi zbadanie wpływu liczby etykiet na jakość generowanej gramatyki

► Szczegóły odwzorowania

Liczba etykiet w funkcji odwzorowania

Język	System etykiet/odwzorowanie	efektywna liczba etykiet w badanym korpusie	maksymalna liczba etykiet
EN	CLAWS-5	-	76
PL	Korpus IPI/map-flexeme-simple	17	19
PL	Korpus IPI/map-flexeme-one	30	32
PL	Korpus IPI/map-gi-flexeme-simple	31	32
PL	Korpus IPI/map-gi-flexeme-one	44	45
PL	Korpus IPI/map-gi-flexeme-simple-eos	60	64
PL	Korpus IPI/map-gi-flexeme-one-eos	83	87
PL	Korpus IPI/map-ipi-one	778	2953
PL	Korpus IPI/map-ipi-one-eos	955	2985
PL	Korpus IPI/map-one	982	2998

Odzworowanie informacji morfologicznej

Przykładowy fragment odzworowania map-gi-flexeme-simple

	subst	depr	num	numcol	adj	adja	adjp	adv	ppron12	ppron3	siebie	...
subst	■											...
depr	■											...
num			■									...
numcol				■								...
adj					■							...
adja					■							...
adjp					■							...
adv								■				...
ppron12									■			...
ppron3										■		...
siebie											■	...
...	...	...	...	...	...	...	...	...	...	...	...	...

Plan prezentacji

- 1 Aktualny stan wiedzy
 - Twierdzenie Golda
 - Pojęcia podstawowe
 - Prace Y. Sakakibary
 - Przetwarzanie wstępne tekstu
- 2 Nowe metody
 - Wyniki badań D. Magermana
 - Wyniki badań M. Gołębskiego
- 3 Rezultaty eksperymentów
- 4 Podsumowanie
 - Wyniki przeprowadzonych prac
 - Zastosowanie
 - Dalsze kierunki badań

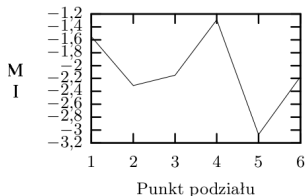


Wyniki badań D. Magermana (rok 1990)

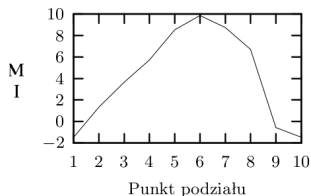
Informacja wzajemna

$$MI(S_1, S_2) = \log \frac{\mathcal{P}(S)}{\mathcal{P}(S_1)\mathcal{P}(S_2)}$$

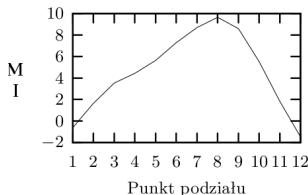
Wyniki badań D. Magermana (rok 1990)



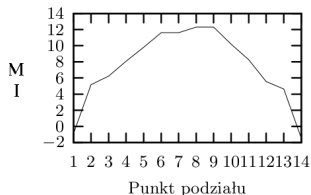
Dwa dni temu był tu Zbyszko .



- To mistrz ujmie się za nami i będzie wojna .



Tak też uczynię , jak radzisz , żeby m tak zdrów był !



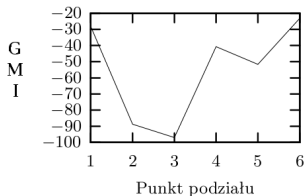
Cała jego postawa zdradzała człowieka dumnego , przywykłego do rozkazywania i ufego w siebie .

Wyniki badań D. Magermana (rok 1990)

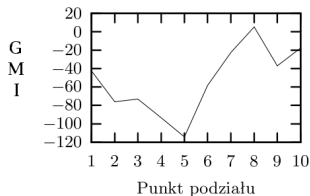
Uogólniona informacja wzajemna

$$\mathcal{GMI}(x_1, \dots, x_i, y_1, \dots, y_i) = \sum_{\substack{X \text{ zakończone } x_i \\ Y \text{ rozpoczęte } y_1}} \frac{1}{\sigma_{XY}} \mathcal{MI}(X, Y)$$

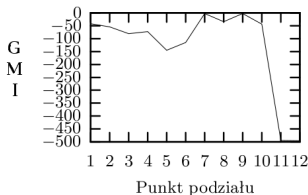
Wyniki badań D. Magermana (rok 1990)



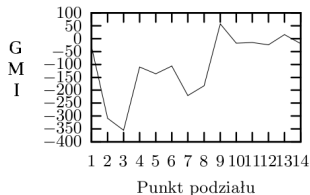
Dwa dni temu był tu Zbyszko .



- To mistrz ujmie się za nami i będzie wojna .



Tak też uczynię , jak radzisz , żeby m tak zdrów był !



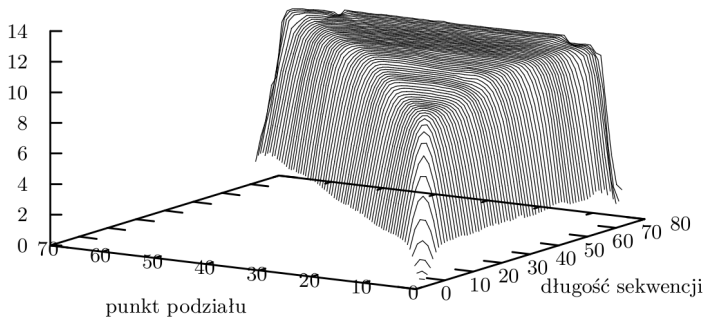
Cała jego postawa zdradzała człowieka dumnego , przywykłego do rozkazywania i ufego w siebie .



Badania wstępne

Wartość średnia MI

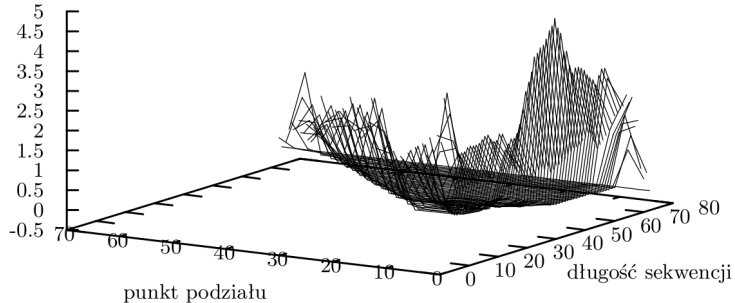
wartość średnia MI



Badania wstępne

Wartość wariancji MI

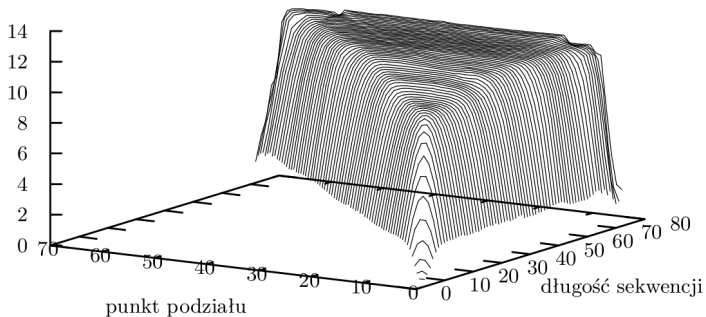
wartość wariancji MI



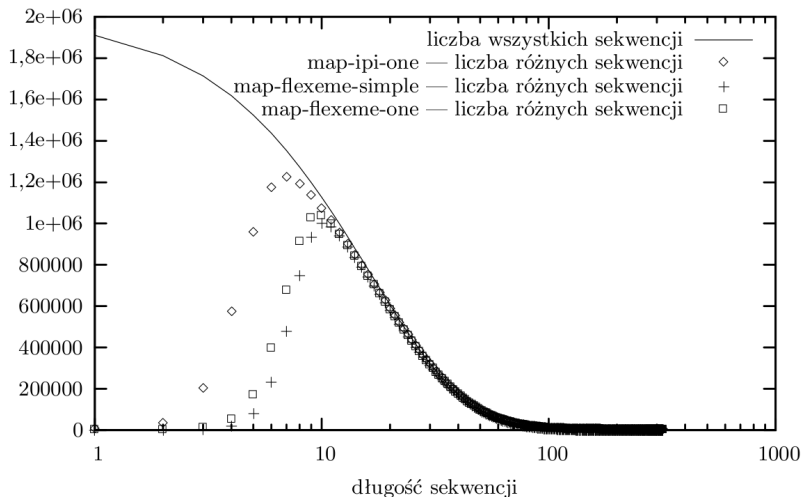
Badania wstępne

Wartość średnia MI

wartość średnia MI



Liczba wystąpień podsekwencji w funkcji długości



Plan prezentacji

- 1 Aktualny stan wiedzy
 - Twierdzenie Golda
 - Pojęcia podstawowe
 - Prace Y. Sakakibary
 - Przetwarzanie wstępne tekstu
- 2 Nowe metody
 - Wyniki badań D. Magermana
 - Wyniki badań M. Gołębskiego
- 3 Rezultaty eksperymentów
- 4 Podsumowanie
 - Wyniki przeprowadzonych prac
 - Zastosowanie
 - Dalsze kierunki badań



Kryteria oceny podziału

Kryterium MI

$$MI(x, y) = \log \frac{\mathcal{P}_{XY}(xy)}{\mathcal{P}_X(x)\mathcal{P}_Y(y)}$$

Kryteria oceny podziału

Kryterium $\mathcal{G}_{IMI}$

$$\mathcal{G}_{IMI}(x, \dots, y) = \log \frac{\mathcal{P}_{X\dots Y}(x, \dots, y)}{\mathcal{P}_X(x) \dots \mathcal{P}_Y(y)}$$



Kryteria oceny podziału

Kryterium $\mathcal{G}_{\mathcal{I}\mathcal{M}\mathcal{I}}_{\text{Norm}}$

$$\mathcal{G}_{\mathcal{I}\mathcal{M}\mathcal{I}}_{\text{Norm}}(x, \dots, y) = \log \frac{\frac{\mathcal{P}_{X\dots Y}(x, \dots, y)}{N_{|X\dots Y|}}}{\frac{\mathcal{P}_X(x)}{N_{|X|}} \dots \frac{\mathcal{P}_Y(y)}{N_{|Y|}}}$$

N_1 — współczynnik dany wzorem:

$$N_1 = \sum_{x \in X_1} \mathcal{P}_{X_1}^2(x)$$

Kryteria oceny podziału

Kryterium $\mathcal{SF}$

$$\mathcal{SF}(x, \dots, y) = \log \frac{1}{\frac{p_X(x)}{N_{|X|}} \dots \frac{p_Y(y)}{N_{|Y|}}}$$

N_1 — współczynnik dany wzorem:

$$N_1 = \sum_{x \in X_1} p_{X_1}^2(x)$$

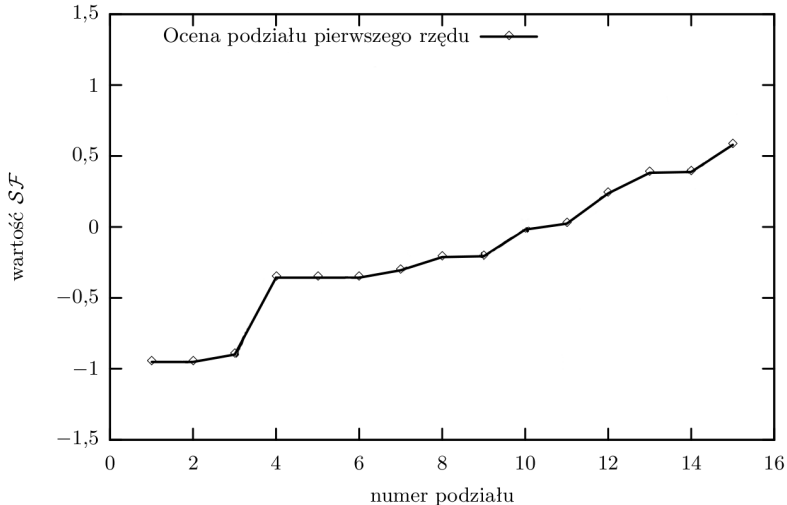
Przykład

Przykładowe zdanie

Duży pies goni małą dziewczynkę



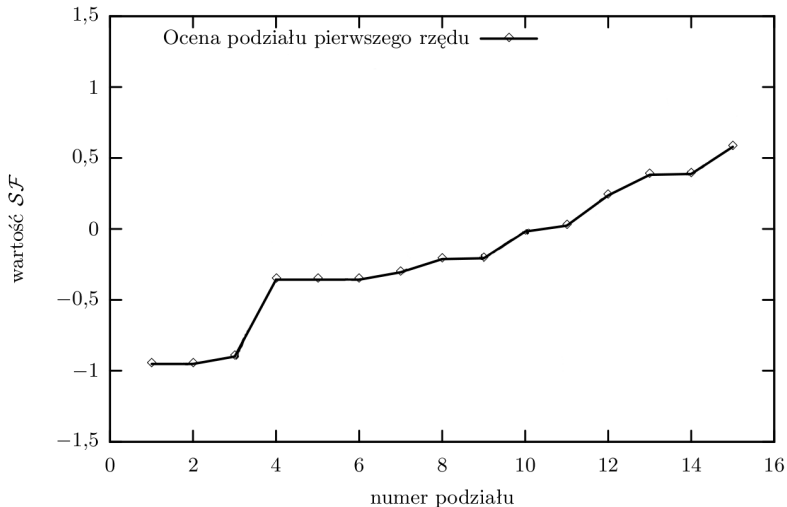
Ocena podziału pierwszego rzędu



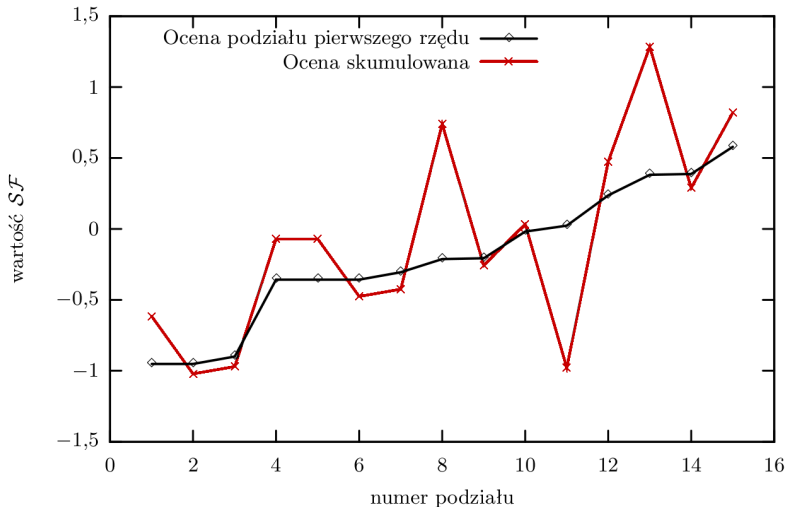
Podziały pierwszego rzędu

Nr	Ocena	Podział
1.	-0.95123	<Duży pies><goni><małą dziewczynkę>
2.	-0.95036	<Duży pies><goni małą dziewczynkę>
⋮	⋮	⋮
11.	0.02321	<Duży><pies goni małą dziewczynkę>
⋮	⋮	⋮

Ocena szkieletu



Ocena szkieletu

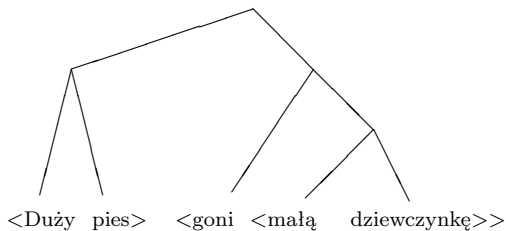


Wybrane szkielety

Ocena addytywna szkieletów

Nr	Ocena	Podział
2.	-1.02120	<<Duży><pies>><<goni><<małą><dziewczynkę>>>
11.	-0.97754	<Duży><<pies><<goni><<małą><dziewczynkę>>>>
⋮	⋮	⋮
1.	-0.61665	<<Duży><pies>><goni><<małą><dziewczynkę>>
⋮	⋮	⋮

Wygenerowana struktura szkieletowa



Kryterium podziału

Eksperymenty potwierdziły, że możliwe jest sformułowanie statystycznego kryterium oceny wielopunktowych podziałów zdań, które nie preferuje konkretnej długości podsekwencji.



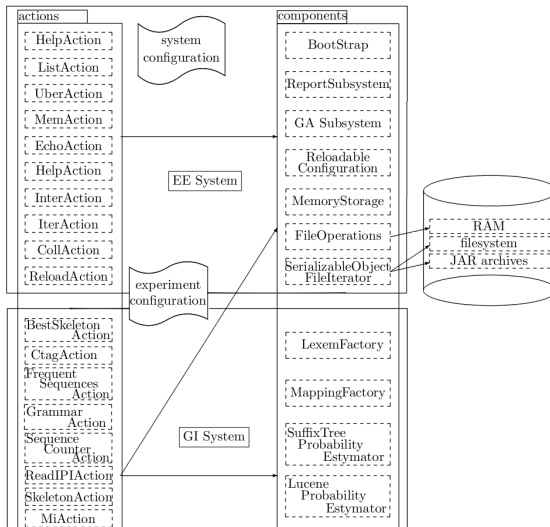
POLITECHNIKA
WARSZAWSKA



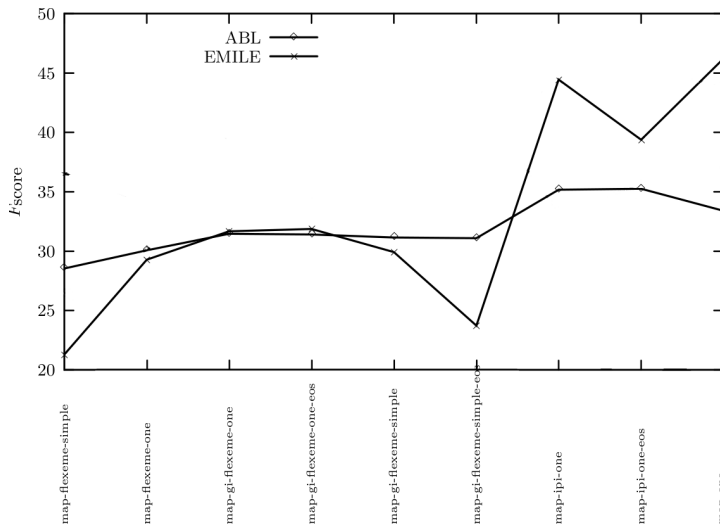
System eksperymentalny

- zbudowano system eksperymentalny rozwiązujący podstawowe problemy jakie pojawiają się podczas wykonywania eksperymentów: dostęp do plików, raporty, obsługa sytuacji wyjątkowych, batch mode vs. interakcja z użytkownikiem, skrypty, wielowątkowość, etc.
- zaimplementowano kryteria:
 - MI
 - $\mathcal{G}MI$
 - $\mathcal{G}_I MI$
 - $\mathcal{G}_I MI_{\text{Norm}}$
 - SF
- zaimplementowano algorytmy (m.in.):
 - wykrywania struktury zdania
 - generowania najlepszych podziałów
 - generowania reguł gramatyki RCFG

Schemat systemu eksperymentalnego



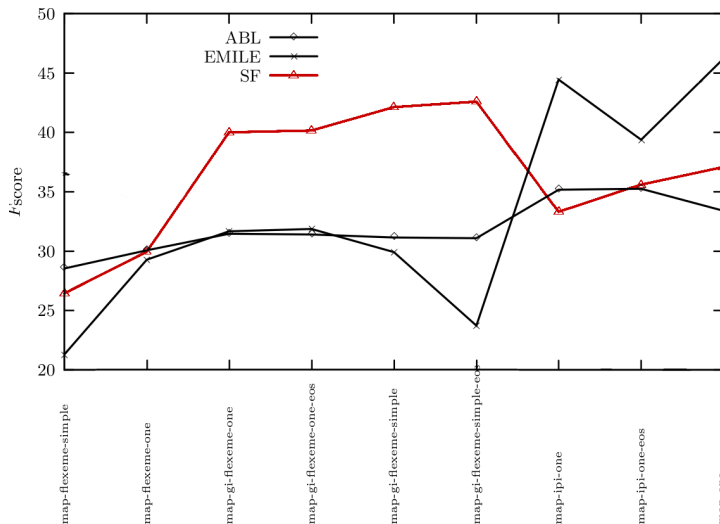
Porównanie $\mathcal{SF}$ z ABL i EMILE



POLITECHNIKA
WARSZAWSKA



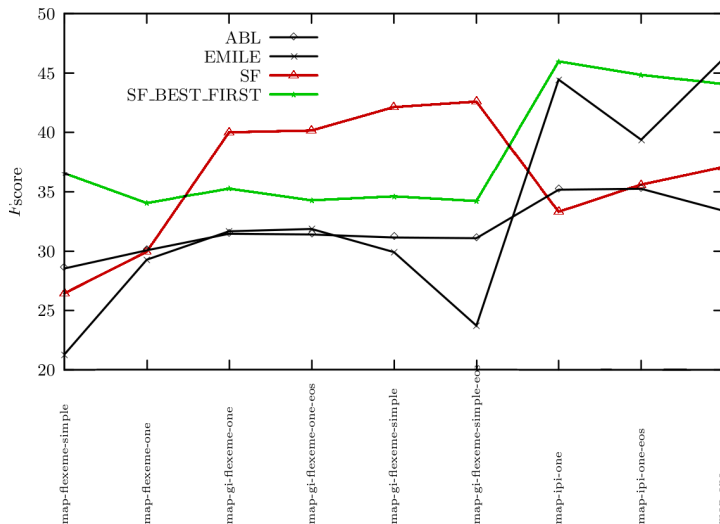
Porównanie $\mathcal{SF}$ z ABL i EMILE



POLITECHNIKA
WARSZAWSKA



Porównanie $\mathcal{SF}$ z ABL i EMILE



POLITECHNIKA
WARSZAWSKA

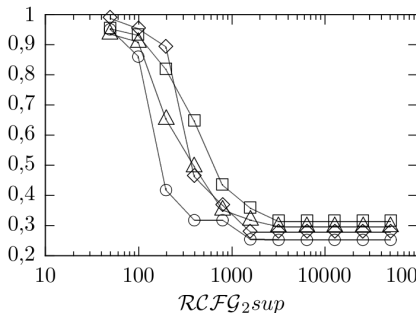


Skuteczność kryterium (struktury szkieletowe)

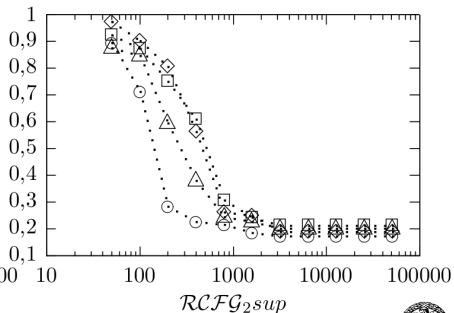
Eksperymenty potwierdzają skuteczność tego kryterium do oceny jakości struktur szkieletowych przy budowie anotowanych strukturalnie korpusów języka naturalnego.

Pokrycie korpusu „15” i „30”

Pokrycie dla korpusu-15:

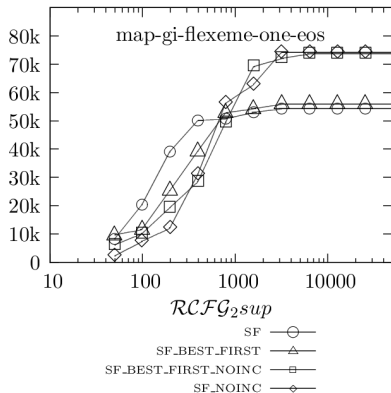


Pokrycie dla korpusu-30:

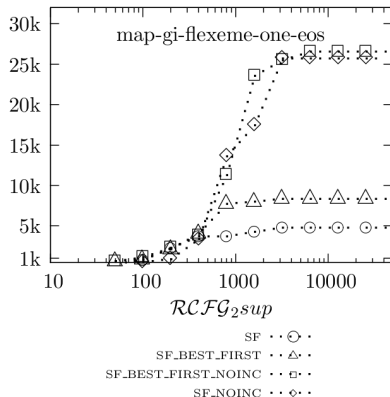


Liczba reguł i symboli nieterminalnych

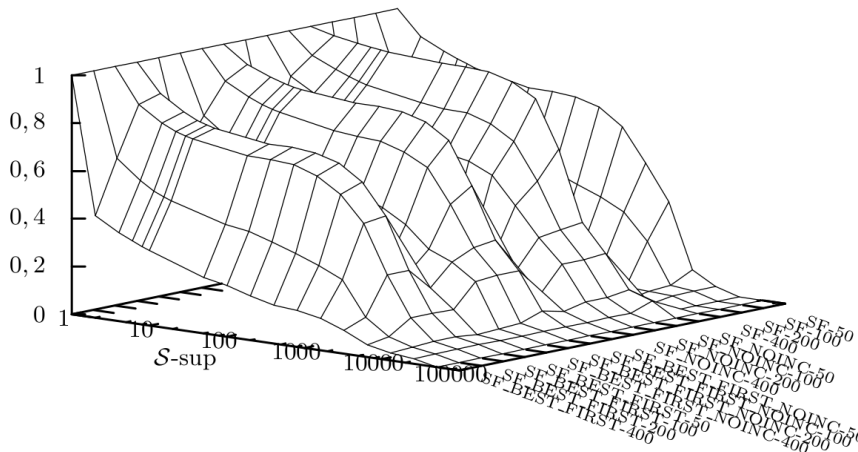
liczba reguł



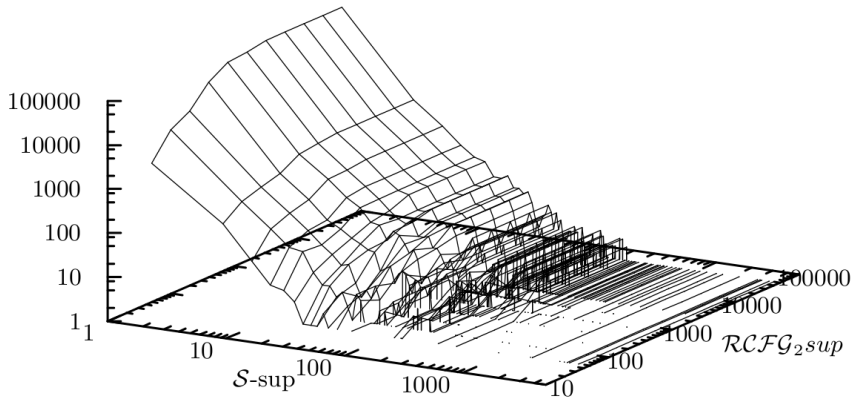
liczba symboli nieterminalnych



Pokrycie zbioru reguł

algorytm- $\mathcal{RCFG}_2$

Liczba reguł



Plan prezentacji

- 1 Aktualny stan wiedzy
 - Twierdzenie Golda
 - Pojęcia podstawowe
 - Prace Y. Sakakibary
 - Przetwarzanie wstępne tekstu
- 2 Nowe metody
 - Wyniki badań D. Magermana
 - Wyniki badań M. Gołębskiego
- 3 Rezultaty eksperymentów
- 4 Podsumowanie
 - Wyniki przeprowadzonych prac
 - Zastosowanie
 - Dalsze kierunki badań



Podsumowanie

Wyniki przeprowadzonych prac

- wykryto zjawiska zaniku różnorodności i gwałtownego spadku ogólnej liczby sekwencji ze wzrostem ich długości
- opracowano metodę oceny podziałów wielopunktowych niewrażliwą na długość sekwencji — kryterium $\mathcal{SF}$
- opracowano metodę oceny struktur szkieletowych na podstawie oceny podziałów tę strukturę definiujących
- opracowano metodę zawężania i poszerzania informacji morfologicznej
- opracowano efektywny algorytm generowania struktur szkieletowych
- udoskonalono metody generowania gramatyki RCFG

Plan prezentacji

- 1 Aktualny stan wiedzy
 - Twierdzenie Golda
 - Pojęcia podstawowe
 - Prace Y. Sakakibary
 - Przetwarzanie wstępne tekstu
- 2 Nowe metody
 - Wyniki badań D. Magermana
 - Wyniki badań M. Gołębskiego
- 3 Rezultaty eksperymentów
- 4 Podsumowanie
 - Wyniki przeprowadzonych prac
 - **Zastosowanie**
 - Dalsze kierunki badań

Podsumowanie

Zastosowanie

- identyfikacja autora (wykrywanie plagiatów)
- wykrywanie błędów w korpusach anotowanych strukturalnie
- wykrywanie zjawisk w języku
- budowa efektywnych parserów
- Semantic Web



Plan prezentacji

- 1 Aktualny stan wiedzy
 - Twierdzenie Golda
 - Pojęcia podstawowe
 - Prace Y. Sakakibary
 - Przetwarzanie wstępne tekstu
- 2 Nowe metody
 - Wyniki badań D. Magermana
 - Wyniki badań M. Gołębskiego
- 3 Rezultaty eksperymentów
- 4 Podsumowanie
 - Wyniki przeprowadzonych prac
 - Zastosowanie
 - Dalsze kierunki badań



Podsumowanie

Dalsze kierunki badań

- metody analizy otrzymanych reguł gramatyki
- metody automatycznego nazywania wykrytych kategorii syntaktycznych
- efektywne estymowanie prawdopodobieństw występowania sekwencji w korpusie



Dziękuję za uwagę



Informacja morfologiczna

Kolejność etykiet użyta w odwzorowaniu

nr	etykieta	nr	etykieta	nr	etykieta	nr	etykieta	nr	etykieta
1.	subst	18.	inf	35.	nom	52.	sup	69.	lbrt
2.	depr	19.	pcon	36.	gen	53.	imperf	70.	rbrt
3.	num	20.	pant	37.	dat	54.	perf	71.	comma
4.	numcol	21.	ger	38.	acc	55.	aff	72.	fstp
5.	adj	22.	pact	39.	inst	56.	neg	73.	ast
6.	adja	23.	ppas	40.	loc	57.	akc	74.	pls
7.	adjp	24.	winien	41.	voc	58.	nakc	75.	mns
8.	adv	25.	pred	42.	m1	59.	praep	76.	cln
9.	ppron12	26.	prep	43.	m2	60.	npraep	77.	scln
10.	ppron3	27.	conj	44.	m3	61.	congr	78.	qst
11.	siebie	28.	qub	45.	f	62.	rec	79.	sie
12.	fin	29.	xxs	46.	n	63.	nagl	80.	nie
13.	bedzie	30.	xxx	47.	pri	64.	agl	81.	eos
14.	aglt	31.	ign	48.	sec	65.	wok		
15.	praet	32.	interp	49.	ter	66.	nwok		
16.	impt	33.	sg	50.	pos	67.	excl		
17.	imps	34.	pl	51.	comp	68.	quot		