

Przetwarzanie języka naturalnego

dr inż. Marcin Ciura

Wydział Informatyki i Telekomunikacji Politechniki Krakowskiej

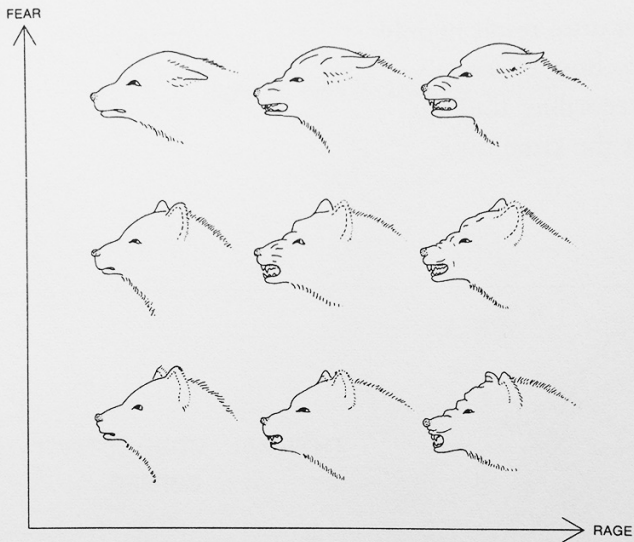
Plan na dziś: 43 slajdy

w tym: 3 filmy, 1 demonstracja i 1 dowcip

- Komunikacja a język
- Gramatyki generatywne
- Analiza składniowa

Komunikacja a język

Komunikacja zwierząt (1)



Komunikacja zwierząt (2)

Sikora bogatka

Komunikacja zwierząt (3)

Pszczoła miodna

Komunikacja ludzi (1)

Bajka Schleichera

Komunikacja ludzi (2): 1878 r., sklep Wokulskiego

— Kaloszyków żąda szanowny pan? Który numer, jeżeli wolno spytać? Ach, szanowny pan zapewne nie pamięta! nie każdy ma czas myśleć o numerze swoich kaloszy, to należy do nas. Szanowny pan pozwoli, że przymierzemy?... Szanowny pan raczy zająć miejsce na taburecie. Paweł! przynieś ręcznik, zdejść panu kalosze i wytrzyj obuwie...



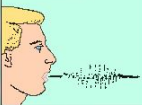
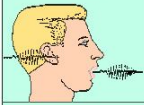
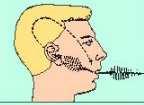


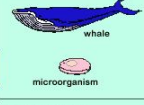
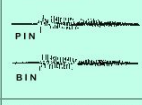




— Bardzo prosimy — mówił prędko Mraczewski — to nasz obowiązek. Zdaje mi się, że te będą dobre — ciągnął, podając parę zczepionych nitką kaloszy. — Doskonałe, pysznie wyglądają; szanowny pan ma tak normalną nogę, że nie podobna mylić się co do numeru. Szanowny pan życzy sobie zapewne literki; jakie mają być literki?...

— Cztery pięćdziesiąt. Potwierdzenie?

Komunikacja ludzi (4): z kolekcji Donalda Knutha



Cechy języków naturalnych wg Hocketta (1960)

<p>1. Vocal-Auditory Channel</p> 	<p>2. Broadcast Transmission and directional reception</p> 	<p>3. Rapid Fading</p> 
<p>4. Interchangeability</p> 	<p>5. Total Feedback</p> 	<p>6. Specialization</p> 
<p>7. Semantics</p>  <p>Pass the salt</p>	<p>8. Arbitrariness</p>  <p>whale</p> <p>microorganism</p>	<p>9. Discreteness</p>  <p>P I N</p> <p>B I N</p>
<p>10. Displacement</p>  <p>Shades of Julius Caesar!</p>	<p>11. Productivity</p>  <p>She has purple hair!</p>	<p>12. Traditional Transmission</p>  <p>What's that?</p> <p>That's an igloo.</p>
<p>13. Duality of Patterning</p>  <p>t.....m</p> <p>m.....t</p>		

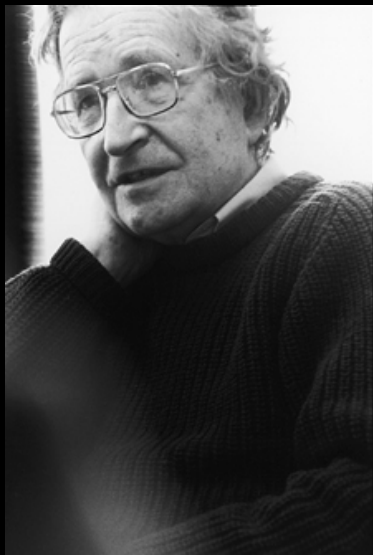
Poddziedziny przetwarzania języka naturalnego

- Fonologia
- Morfologia
- Składnia
- Semantyka
- Pragmatyka

Gramatyki generatywne

Noam Chomsky

Syntactic Structures (1957)



Gramatyka generatywna: przykład

$S \rightarrow NP VP$

$NP \rightarrow Det N$

$NP \rightarrow Det Adj N$

$VP \rightarrow V$

$VP \rightarrow V NP$

$Det \rightarrow a$

$Det \rightarrow the$

$N \rightarrow cat$

$N \rightarrow mouse$

$Adj \rightarrow black$

$Adj \rightarrow white$

$V \rightarrow ate$

$V \rightarrow saw$

Notacja Backusa-Naura (BNF)

$S ::= NP VP$

$NP ::= Det N \mid Det Adj N$

$VP ::= V \mid V NP$

$Det ::= a \mid the$

$N ::= cat \mid mouse$

$Adj ::= black \mid white$

$V ::= ate \mid saw$

Definicja gramatyki generatywnej (1)

Cztery składniki gramatyki generatywnej:

1. Skończony zbiór *symboli nieterminalnych*
2. Skończony zbiór *symboli terminalnych*
3. Skończony zbiór *produkcji*
4. Wyróżniony nieterminalny *symbol startowy*

$S ::= NP VP$

$Det ::= a \mid the$

$NP ::= Det N \mid Det Adj N$

$N ::= cat \mid mouse$

$VP ::= V \mid V NP$

$Adj ::= black \mid white$

$V ::= ate \mid saw$

Definicja gramatyki generatywnej (2)

Cztery składniki gramatyki generatywnej:

1. Skończony zbiór *symboli nieterminalnych*: S, NP, VP, Det, N, A, V
2. Skończony zbiór *symboli terminalnych*
3. Skończony zbiór *produkcji*
4. Wyróżniony nieterminalny *symbol startowy*

$$S ::= NP VP$$
$$Det ::= a \mid the$$
$$NP ::= Det N \mid Det Adj N$$
$$N ::= cat \mid mouse$$
$$VP ::= V \mid V NP$$
$$Adj ::= black \mid white$$
$$V ::= ate \mid saw$$

Definicja gramatyki generatywnej (3)

Cztery składniki gramatyki generatywnej:

1. Skończony zbiór *symboli nieterminalnych* S, NP, VP, Det, N, A, V
2. Skończony zbiór *symboli terminalnych*: **a**, **the**, **cat**, **mouse**, ...
3. Skończony zbiór *produkcji*
4. Wyróżniony nieterminalny *symbol startowy*

$S ::= NP VP$

$Det ::= a \mid the$

$NP ::= Det N \mid Det Adj N$

$N ::= cat \mid mouse$

$VP ::= V \mid V NP$

$Adj ::= black \mid white$

$V ::= ate \mid saw$

Definicja gramatyki generatywnej (4)

Cztery składniki gramatyki generatywnej:

1. Skończony zbiór *symboli nieterminalnych* S, NP, VP, Det, N, A, V
2. Skończony zbiór *symboli terminalnych*: a, the, cat, mouse, ...
3. Skończony zbiór *produkcji*
4. Wyróżniony nieterminalny *symbol startowy*

$S ::= NP VP$

$Det ::= a \mid the$

$NP ::= Det N \mid Det Adj N$

$N ::= cat \mid mouse$

$VP ::= V \mid V NP$

$Adj ::= black \mid white$

$V ::= ate \mid saw$

Definicja gramatyki generatywnej (5)

Cztery składniki gramatyki generatywnej:

1. Skończony zbiór *symboli nieterminalnych* S, NP, VP, Det, N, A, V
2. Skończony zbiór *symboli terminalnych*: a, the, cat, mouse, ...
3. Skończony zbiór *produkcji*
4. Wyróżniony nieterminalny *symbol startowy*: S

$S ::= NP VP$

$Det ::= a \mid the$

$NP ::= Det N \mid Det Adj N$

$N ::= cat \mid mouse$

$VP ::= V \mid V NP$

$Adj ::= black \mid white$

$V ::= ate \mid saw$

Definicja gramatyki generatywnej (6)

Cztery składniki gramatyki generatywnej:

1. Skończony zbiór *symboli nieterminalnych* S, NP, VP, Det, N, A, V
2. Skończony zbiór *symboli terminalnych*: a, the, cat, mouse, ...
3. Skończony zbiór *produkcji*
4. Wyróżniony nieterminalny *symbol startowy*: S

$S ::= NP VP$

$Det ::= a \mid the$

$NP ::= Det N \mid Det Adj N$

$N ::= cat \mid mouse$

$VP ::= V \mid V NP$

$Adj ::= black \mid white$

$V ::= ate \mid saw$

Hierarchia Chomsky'ego

- typ 3 Gramatyki regularne
- typ 2 Gramatyki bezkontekstowe
- typ 1 Gramatyki kontekstowe
- typ 0 Gramatyki bez ograniczeń
(rekurencyjnie przeliczalne)

Gramatyki regularne (1)

A, B, \dots : symbole nieterminalne

a, b, \dots : symbole terminalne

Wszystkie produkcje w gramatykach *prawostronnie regularnych* mają jedną z trzech postaci:

$$A \rightarrow a$$

$$A \rightarrow aB$$

$$A \rightarrow \varepsilon$$

Gramatyki regularne (2)

A, B, \dots : symbole nieterminalne

a, b, \dots : symbole terminalne

Wszystkie produkcje w gramatykach *lewostronnie regularnych* mają jedną z trzech postaci:

$$A \rightarrow a$$

$$A \rightarrow Ba$$

$$A \rightarrow \varepsilon$$

Gramatyki regularne (3)

1. Każda gramatyka regularna (*regular grammar*) musi być albo lewostronnie regularna, albo prawostronnie regularna
2. Każda gramatyka regularna generuje jakiś *język regularny*
3. Każdej gramatyce regularnej odpowiadają jakieś automaty skończone

Gramatyki bezkontekstowe (1)

A, B, \dots : symbole nieterminalne

a, b, \dots : symbole terminalne

Wszystkie produkcje w gramatyce bezkontekstowej (*context-free grammar*) mają postać

$$A \rightarrow \alpha,$$

gdzie

- α oznacza dowolny skończony ciąg symboli terminalnych lub nieterminalnych,
np. ε , cat, V , the N , Adj N itp.
- A oznacza dowolny symbol nieterminalny,
np. S , VP , Det itp.

Gramatyki bezkontekstowe (2)

1. Każda gramatyka bezkontekstowa generuje jakiś *język bezkontekstowy*
2. Każdej gramatyce bezkontekstowej odpowiadają jakieś *automaty ze stosem*

Gramatyki kontekstowe (1)

A, B, \dots : symbole nieterminalne

a, b, \dots : symbole terminalne

Wszystkie produkcje w gramatyce kontekstowej (*context-sensitive grammar*) mają postać

$$\alpha A \beta \rightarrow \alpha \gamma \beta,$$

gdzie

- α i β oznacza dowolny skończony (być może pusty) ciąg symboli terminalnych lub nieterminalnych,
- γ oznacza dowolny skończony niepusty ciąg symboli terminalnych lub nieterminalnych,
- A oznacza dowolny symbol nieterminalny

Gramatyki kontekstowe (2)

1. Każda gramatyka kontekstowa generuje jakiś *język kontekstowy*
2. Każdej gramatyce kontekstowej odpowiadają jakieś *automaty liniowo ograniczone*

Gramatyki bez ograniczeń (1)

A, B, \dots : symbole nieterminalne

a, b, \dots : symbole terminalne

Wszystkie produkcje w gramatyce bez ograniczeń (*unrestricted grammar*) mają postać

$$\alpha \rightarrow \beta,$$

gdzie

- α oznacza dowolny skończony i niepusty ciąg symboli terminalnych lub nieterminalnych,
- β oznacza dowolny skończony (być może pusty) ciąg symboli terminalnych lub nieterminalnych

Gramatyki bez ograniczeń (2)

1. Każda gramatyka bez ograniczeń generuje jakiś *język rekurencyjnie przeliczalny*
2. Każdej gramatyce bez ograniczeń odpowiadają jakieś *maszyny Turinga*

Do której klasy należą języki naturalne?

Konsensus: raczej bezkontekstowe.

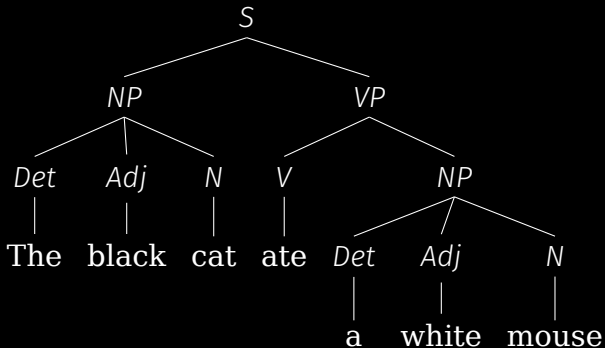
Udowodnione wyjątki:

1. szwajcarski dialekt języka niemieckiego (Shieber, 1985)
2. bambara, używany w Afryce Zachodniej (Culy, 1985)
3. szwedzki (Miller, 1991)
4. angielski (Higginbotham, 1987)
5. rumuński (Longenbaugh, 2011)

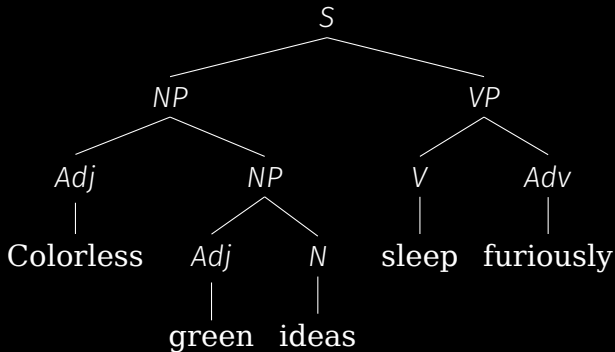
Analiza składniowa

- Drzewa rozbioru
- Niejednoznaczności (*ambiguities*)
- Analizatory składniowe (parsery)

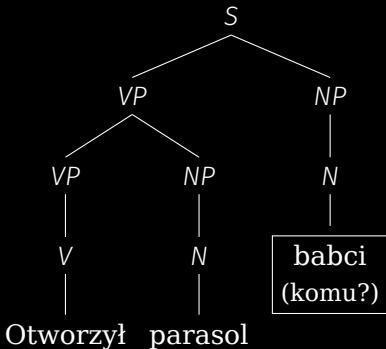
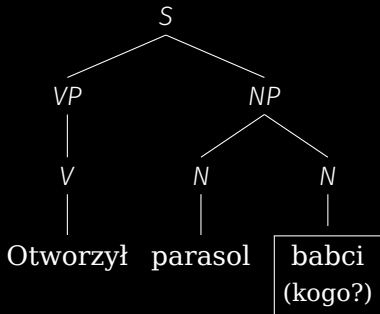
Drzewa rozbioru (1)



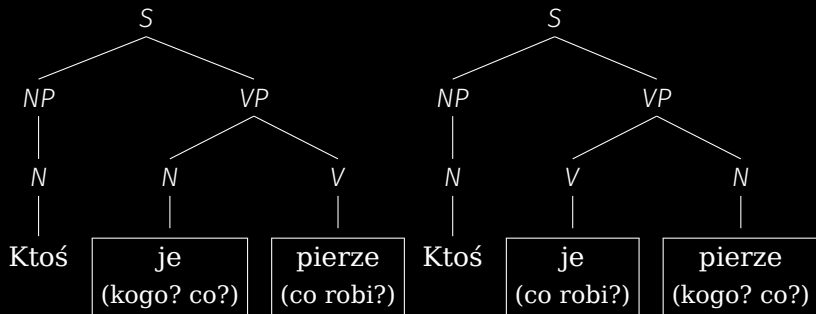
Drzewa rozbioru (2)



Niejednoznaczności w języku (1)



Niejednoznaczności w języku (2)



Analizatory składniowe (parsery)

- Dla gramatyk regularnych: automaty skończone
- Dla gramatyk bezkontekstowych — mnogość algorytmów:
 - dla jednoznacznych (*unambiguous*) gramatyk bezkontekstowych np. LR(1) — analiza wstępująca (*bottom-up parsing*), LL(1) — analiza zstępująca (*top-down parsing*); złożoność zwykle $O(N)$
 - dla niejednoznacznych (*ambiguous*) gramatyk bezkontekstowych np. algorytm CYK (Cocke-Younger-Kasami) — analiza wstępująca, algorytm Earleya — hybryda; złożoność pesymistyczna zwykle $O(N^3)$; złożoność średnia może być $O(N)$
 - generatory analizatorów leksykalnych
- Dla gramatyk typu 2 i 3 — praktycznie nic

Wpływ Chomsky'ego w językach programowania

Ogromny:

- Analizator leksykalny zazwyczaj korzysta z gramatyki regularnej, np.
 $\langle \text{identyfikator} \rangle ::= [A-Za-z_][A-Za-z_0-9]^*$
- Analizator składniowy często korzysta z gramatyki bezkontekstowej, np.
 $\langle \text{instrukcja warunkowa} \rangle ::=$
 $\text{if } (\langle \text{wyrażenie} \rangle) \langle \text{instrukcja} \rangle$

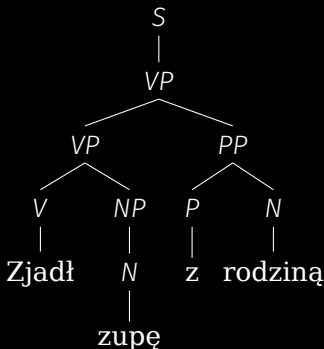
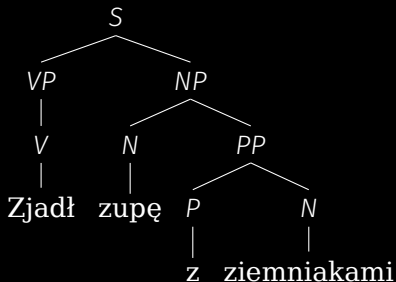
Wpływ Chomsky'ego w językach naturalnych

Taki sobie:

- Przekonanie językoznawców, że gramatyka ma określać, co jest poprawne, a co niepoprawne w języku
- *All grammars leak* (Sapir, 1921)
- Niejednoznaczności

Niejednoznaczności — ciąg dalszy (1)

Rozstrzygane na poziomie semantycznym: np.
przyłączanie fraz przyimkowych (*prepositional phrase attachment, PP attachment*)



I saw a man with a telescope

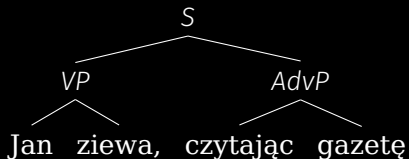
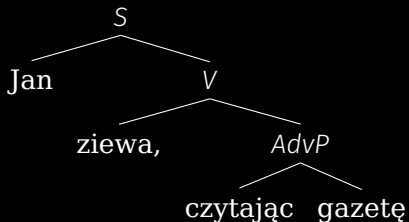
Niejednoznaczności — ciąg dalszy (2)

Rozstrzygane na poziomie pragmatycznym, np.

To jest siostra Basi, którą ci wczoraj przedstawiłem

Niejednoznaczności — ciąg dalszy (3)

Fałszywe niejednoznaczności, występujące tylko w gramatyce, bez odzwierciedlenia w języku, np.



Niejednoznaczności — ciąg dalszy (4)

Przykład: Świgr (gramatyka Świdzińskiego)

Podsumowanie

- Komunikacja a język
- Gramatyki generatywne
 - hierarchia Chomsky'ego
 - gramatyki regularne i bezkontekstowe
- Analiza gramatyczna
 - drzewa rozbioru
 - niejednoznaczności
 - wpływ Chomsky'ego

**Do zobaczenia
na następnym wykładzie
o analizie zależnościowej**
