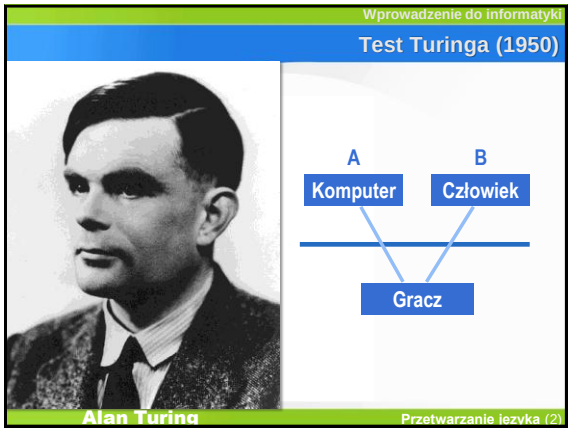
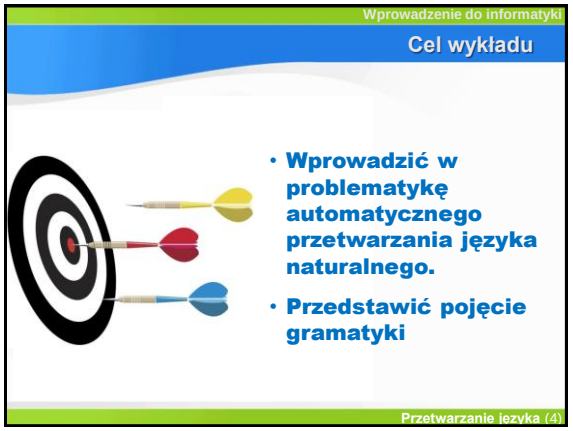
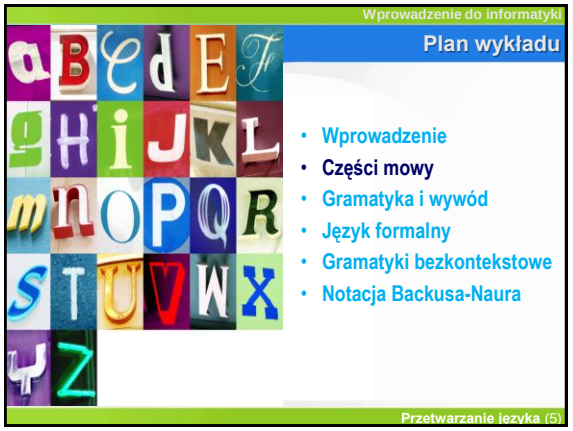
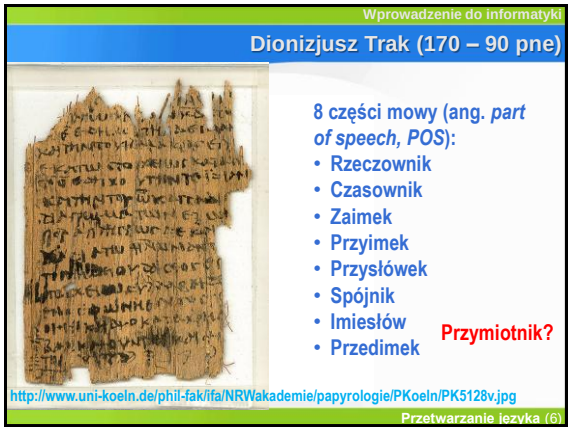

Wprowadzenie do informatyki
Jerzy Nawrocki
Wydział Informatyki
Politechnika Poznańska
jerzy.nawrocki@put.poznan.pl
Sztuczna inteligencja i język naturalny


Wprowadzenie do informatyki
Test Turinga (1950)
A
B
Komputer
Człowiek
Gracz
Alan Turing
Przetwarzanie języka (2)


Wprowadzenie do informatyki
Program ELIZA (1966)
User: Men are all alike.
Eliza: In what way?
User: They're always bugging us about something or other.
Eliza: Can you think of a specific example?
User: My boyfriend made me come here.
Eliza: Your boyfriend made you come here?
Joseph Weizenbaum
Przetwarzanie języka (3)


Wprowadzenie do informatyki
Cel wykładu
Wprowadzić w problematykę automatycznego przetwarzania języka naturalnego.
Przedstawić pojęcie gramatyki
Przetwarzanie języka (4)


Wprowadzenie do informatyki
Plan wykładu
Wprowadzenie
Części mowy
Gramatyka i wywód
Język formalny
Gramatyki bezkontekstowe
Notacja Backusa-Naura
Przetwarzanie języka (5)


Wprowadzenie do informatyki
Dionizjusz Trak (170 – 90 pne)
8 części mowy (ang. part of speech, POS):
Rzeczownik
Czasownik
Zaimek
Przyimek
Przysłówek
Spójnik
Imiesłów
Przedimek
Przymiotnik?
http://www.uni-koeln.de/phil-fak/ifa/NRWakademie/papyrologie/PKoeln/PK5128v.jpg
Przetwarzanie języka (6)

Współczesna systematyka części mowy – język ang.

- Rzeczownik
- Czasownik
- Zaimek
- Przyimek
- Przysłówek
- Spójnik
- Imiesłów
- Przedimek

Część mowy ↔ znacznik

Penn Treebank: 45 znaczników
Znaczniki C5: 61 znaczników
Brown corpus: 87 znaczników
Znaczniki C7: 146 znaczników

(znacznik = ang. *tag*)

Znaczniki Penn Treebank

CC Coordinating conjunction *and*
CD Cardinal number *one*, *two*
DT Determiner *the*
EX Existential *there* ***there*** *are*
FW Foreign word *mea culpa*
IN Preposition or subordinating conjunction *of*, *in*, *by*
JJ Adjective *yellow*
JJR Adjective, comparative *bigger*
NN Noun, singular or mass *tiger*
NNS Noun, plural *tigers*
VB Verb, base form *eat*
VBD Verb, past tense *ate*

Oznaczanie części mowy

The grand jury commented on a number of other topics.

The/*DT* grand/*JJ* jury/*NN* commented/*VBD* on/*IN* a/*DT* number/*NN* of/*IN* other/*JJ* topics/*NNS* ./.

(oznaczanie cz. mowy = ang. *tagging*)

Oznaczanie części mowy

Oznaczanie części mowy – Główny problem

Book that flight.

Book/VB that/DT flight/NN ./.

Inne możliwości:
Book/NN *Buy me that book.*
that/CC *I thought that you knew.*

Niejednoznaczność

Rozstrzyganie niejednoznaczności

Book that flight.

Book/VB/NN that/DT/CC flight/NN ./.

Rozstrzyganie niejednoznaczności

Book that flight.

Book/VB/NN that/DT/CC flight/NN ./.

Książka ten/DT lot.
Książka, że/CC lot.

Reguła:

Jeśli jest konflikt między czasownikiem (VB, VBD, ..) a inną częścią mowy i w zdaniu nie ma innego czasownika, to należy uznać, że to jest czasownik.

Rozstrzyganie niejednoznaczności

Book that flight.

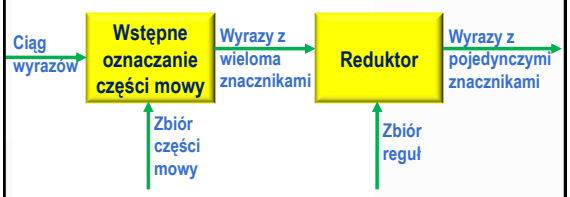
Book/VB/NN that/DT/CC flight/NN ./.

Zarezerwuj ten/DT lot.
Zarezerwuj, że/CC lot.

Reguła:

Jeśli jest konflikt między „że/CC" a inną częścią mowy i w zdaniu jest tylko jeden czasownik, to wariant „że/CC" należy odrzucić.

Regułowe oznaczanie części mowy



System ENGTWOL: około 1100 reguł

Eliminowanie niejednoznaczności

- **Podejście regułowe**
- **Podejście stochastyczne** (ukryte modele Markowa – HMM; pewne części mowy występują częściej niż inne i widać to jeszcze bardziej, gdy uwzględni się kontekst, w którym występuje dany wyraz)

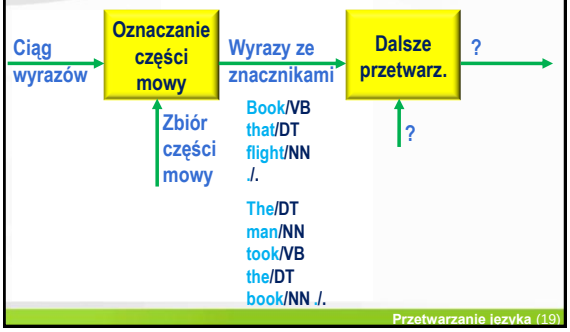
Plan wykładu

Oznaczanie części mowy

Book/VB that/DT flight/NN ./.

The/DT man/NN took/VB the/DT book/NN ./.

Oznaczanie części mowy



Składnia zdania

Noam Chomsky

S → NP VP *We + are smart*

Fraza rzeczownikowa – NP

Noam Chomsky

NP → Pronoun *we*
| ProperNoun *New York*
| Det Nominal *a + flight*
Nominal → Noun *degree*
| Noun Nominal *university degree*

Części mowy: Noun (NN), ProperNoun (NNP), Det (NT), Pronoun
Symbole nieterminalne (części zdania): NP, Nominal

Fraza czasownikowa – VP

Noam Chomsky

VP → Verb *took*
| Verb NP *want + a flight*
Verb → VerbBase *take*
| VerbPast *took*

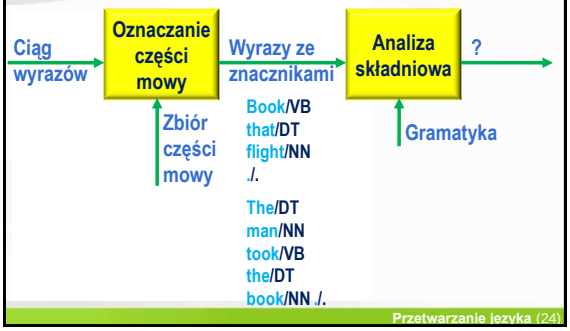
Części mowy: VerbBase (VB), VerbPast (VBD)
Symbole nieterminalne (części zdania): VP, Verb, NP

Gramatyka

Noam Chomsky

S → NP VP
NP → Pronoun
| ProperNoun
| Det Nominal
Nominal → Noun
| Noun Nominal
VP → Verb
| Verb NP
Verb → VerbBase
| VerbPast

Oznaczanie części mowy

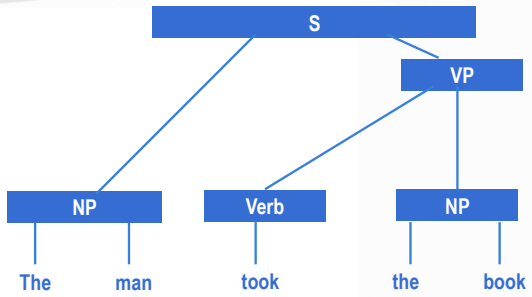


Gramatyka

S → NP VP
NP → Pronoun
| ProperNoun
| Det Nominal
Nominal → Noun
| Noun Nominal
VP → Verb
| Verb NP
Verb → VerbBase
| VerbPast

The/DT
man/NN
took/VB
the/DT
book/NN
./.

Drzewo rozbioru (*parse tree*)



Oznaczanie części mowy

Book/VB
that/DT
flight/NN
./.

The/DT
man/NN
took/VB
the/DT
book/NN ./.

Produkcje i wywód

$1+ = \{1, 11, 111, \ldots\}$

Symbol początkowy: S

Produkcje: 1) $S \to 1$
(reguły zastępowania) 2) $S \to S\,1$

Wywód:
1: $S \overset{1}{\Rightarrow} 1$

111: $S \overset{2}{\Rightarrow} S\,1 \overset{2}{\Rightarrow} S\,1\,1 \overset{1}{\Rightarrow} 1\,1\,1$

Inne produkcje

1) $S \to A\,B$
2) $A \to 1$
3) $A \to A\,1$
4) $B \to 0$
5) $B \to B\,0$

Wywód:

10: $S \overset{1}{\Rightarrow} A\,B \overset{2}{\Rightarrow} 1\,B \overset{4}{\Rightarrow} 1\,0$

100:

Gramatyka

$S \to A\,B$
$A \to 1$
$A \to A\,1$
$B \to 0$
$B \to B\,0$

• Symbol początkowy

Gramatyka

$S \to A\ B$
$A \to 1$
$A \to A\ 1$
$B \to 0$
$B \to B\ 0$

- Symbol początkowy
- Symbole nieterminalne N = {S, A, B}

Gramatyka

$S \to A\ B$
$A \to 1$
$A \to A\ 1$
$B \to 0$
$B \to B\ 0$

- Symbol początkowy
- Symbole nieterminalne N = {S, A, B}
- Symbole terminalne T = {0, 1}

Gramatyka

$S \to A\ B$
$A \to 1$
$A \to A\ 1$
$B \to 0$
$B \to B\ 0$

- Symbol początkowy
- Symbole nieterminalne N = {S, A, B}
- Symbole terminalne T = {0, 1}
- Produkcje

Plan wykładu

- Wprowadzenie
- Części mowy
- Gramatyka i wywód
- Język formalny
- Gramatyki bezkontekstowe
- Notacja Backusa-Naura

Domknięcie relacji wywodu

1) $S \to A\ B$
2) $A \to 1$
3) $A \to A\ 1$
4) $B \to 0$
5) $B \to B\ 0$

Wywód:

$S \overset{1}{\Rightarrow} A\ B \overset{2}{\Rightarrow} 1\ B \overset{4}{\Rightarrow} 1\ 0$

$S \overset{+}{\Rightarrow} 1\ 0$ *Z S można wywieść 10 stosując 1 lub więcej produkcji*

Zbiór ciągów nad alfabetem

$S \to A\ B$
$A \to 1$
$A \to A\ 1$
$B \to 0$
$B \to B\ 0$

Alfabet = Zbiór symboli terminalnych
T = {0, 1}

Zbiór ciągów nad alfabetem T*:

Zbiór wszystkich ciągów skończonych zbudowanych z elementów zbioru T.

Jeśli T = {0, 1} *to* T* = {ε, 0, 1, 00, 01, 10, 11, 000, ...}

Jeśli T = {a, b, c} *to* T* = {ε, a, b, c, aa, ab, ac, ba, bb, bc, ...}

Język formalny

Gramatyka G = <S, N, T, P>

S – Symbol początkowy
N – Zbiór symboli nieterminalnych
T – Zbiór symboli terminalnych
P – Zbiór produkcji

Język formalny L zdefiniowany przez gramatykę G:

$$L(G) = \{x \in T^* : S \overset{+}{\Rightarrow} x\}$$

Język formalny

1) $S \to A\,B$
2) $A \to 1$
3) $A \to A\,1$
4) $B \to 0$
5) $B \to B\,0$

$$L(G) = \{x \in T^* : S \overset{+}{\Rightarrow} x\}$$

$$S \overset{1}{\Rightarrow} A\,B \overset{2}{\Rightarrow} 1\,B$$

$$S \overset{+}{\Rightarrow} 1\,B$$

Czy 1B *należy do* L(G) *?*

$$11 \in T^*$$

Czy 11 *należy do* L(G) *?*

Równoważność gramatyk

Gramatyki G1 *i* G2 *są równoważne wtedy i tylko wtedy, gdy*

$$L(G1) = L(G2)$$

G1	G2	G3
$S \to A\,B$	$S \to S\,0$	$S \to 1\,S$
$A \to 1$	$S \to A\,0$	$S \to 1\,A$
$A \to A\,1$	$A \to 1$	$A \to 0$
$B \to 0$	$A \to A\,1$	$A \to 0\,A$
$B \to B\,0$		

Plan wykładu

- Wprowadzenie
- Części mowy
- Gramatyka i wywód
- Język formalny
- Gramatyki bezkontekstowe
- Notacja Backusa-Naura

Klasyfikacja Chomsky'ego

Noam Chomsky

Klasyfikacja Chomsky'ego

Gramatyki liniowe

a+ b+

Lewoliniowa	Prawoliniowa
1. $S \to S\ b$	1. $S \to a\ S$
2. $S \to A\ b$	2. $S \to a\ B$
3. $A \to a$	3. $B \to b\ B$
4. $A \to A\ a$	4. $B \to b$

Twierdzenie.
Dla każdego wyrażenia regularnego istnieje gramatyka lewoliniowa (prawoliniowa) opisująca ten sam język.

Klasyfikacja Chomsky'ego

Gramatyka bezkontekstowa

1. $W \to (\ W\ )$
2. $W \to 1$

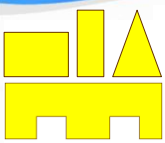
Jeden nieterminal

Gramatyka bezkontekstowa

1. $W \to S$
2. $W \to W + S$
3. $S \to C$
4. $S \to S * C$
5. $C \to L$
6. $C \to (\ W\ )$
7. $L \to 1$
8. $L \to 2$
9. $L \to 3$

Jeden nieterminal

Gramatyki bezkontekstowe potrafią więcej



?

Gramatyki bezkontekstowe potrafią więcej

Język $0^n\ 1^n$, gdzie $n \geq 1$.
0011 *OK.*
0001 *Error*
$S \to 0\ S\ 1$
$S \to 0\ 1$

Język $0^n\ 1^k$, gdzie $n, k \geq 1$.
0001 OK.
1000 Error
$S \to 0\ S$
$S \to 0\ J$
$J \to 1\ J$
$J \to 1$

?

Gramatyki bezkontekstowe potrafią więcej

Język $0^n\,1^n$, gdzie $n \geq 1$.

?

Klasyfikacja Chomsky'ego

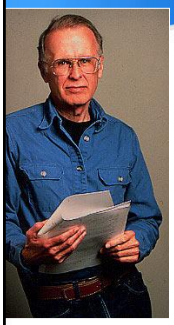
Gramatyka kontekstowa

1. $S \rightarrow a\ X\ Y$
2. $S \rightarrow a\ S\ X\ Y$
3. $a\ X \rightarrow a\ b$
4. $b\ X \rightarrow b\ b$
5. $c\ X \rightarrow c\ c$
6. $b\ Y \rightarrow b\ c$
7. $c\ Y \rightarrow c\ c$

Plan wykładu

- Wprowadzenie
- Części mowy
- Gramatyka i wywód
- Język formalny
- Gramatyki bezkontekstowe
- Notacja Backusa-Naura

Rozszerzona notacja Backusa-Naura



John Backus

Produkcje + wyrażenia regularne

<C> ::= '0' | '1' | '2' | '3' | '4' | '5' | '6' | '7' | '8' | '9'

<L> ::= $\langle C \rangle^+$

<L> ::= $\langle C \rangle^*$ <C>

<S> ::= $(\langle L \rangle\ \text{'*'})^*$ <L>

<W> ::= $(\langle S \rangle\ \text{'+'})^*$ <S>

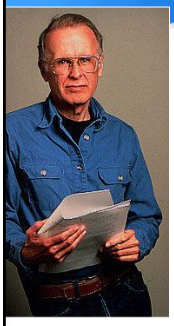
Przejście z EBNF na gramatyki

$\langle J \rangle ::= \langle A \rangle^* \langle B \rangle$

$J \rightarrow B$

$J \rightarrow A\ J$

Przejście z EBNF na gramatyki



<C> ::= ‘0’ | ‘1’ | ‘2’ | ‘3’ | ‘4’ | ‘5’ | ‘6’ | ‘7’ | ‘8’ | ‘9’

<L> ::= <C>* <C>

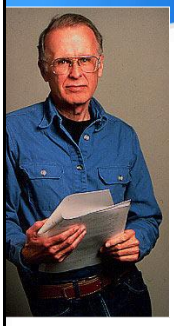
<S> ::= (<L> ‘*’)* <L>

<W> ::= (<S> ‘+’)* <S>

C → ‘0’
C → ‘1’
C → ‘2’
. . .
C → ‘9’

John Backus

Przejście z EBNF na gramatyki



<C> ::= ‘0’ | ‘1’ | ‘2’ | ‘3’ | ‘4’ | ‘5’ | ‘6’ | ‘7’ | ‘8’ | ‘9’

<L> ::= <C>* <C>

<S> ::= (<L> ‘*’)* <L>

<W> ::= (<S> ‘+’)* <S>

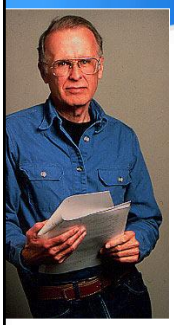
<J> ::= <A>* <B>

J → B
J → A J

L → C
L → C L

John Backus

Przejście z EBNF na gramatyki



<C> ::= ‘0’ | ‘1’ | ‘2’ | ‘3’ | ‘4’ | ‘5’ | ‘6’ | ‘7’ | ‘8’ | ‘9’

<L> ::= <C>* <C>

<S> ::= (<L> ‘*’)* <L>

<W> ::= (<S> ‘+’)* <S>

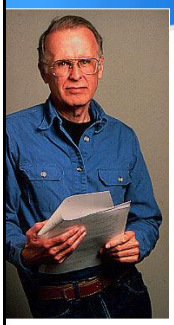
<J> ::= <A>* <B>

J → B
J → A J

S → L
S → L ‘*’ S

John Backus

Przejście z EBNF na gramatyki



<C> ::= ‘0’ | ‘1’ | ‘2’ | ‘3’ | ‘4’ | ‘5’ | ‘6’ | ‘7’ | ‘8’ | ‘9’

<L> ::= <C>* <C>

<S> ::= (<L> ‘*’)* <L>

<W> ::= (<S> ‘+’)* <S>

<J> ::= <A>* <B>

J → B
J → A J

W → S
W → S ‘+’ W

John Backus

Podsumowanie

Przykładowa budowa narzędzia NLP

Tekst
↓ Podział tekstu na zdania
↓ Ciąg zdań
↓ Podział zdania na wyrazy
↓ Ciąg wyrazów
↓ Oznaczanie części mowy
↓ Wyrazy ze znacznikami
↓ Lematyzer
↓ Dochodzi postać podst. wyrazów
↓ Analiza składniowa
↓ Drzewo rozbioru składniowego

Oznaczanie części mowy

The grand jury commented on a number of other topics.

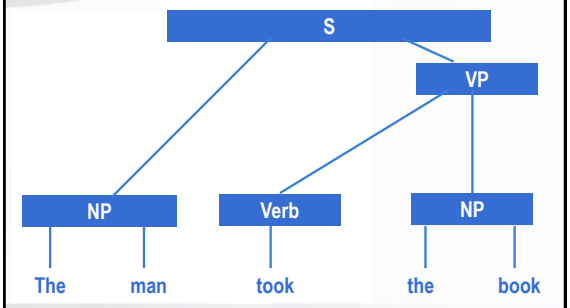
↓

The/*DT* grand/*JJ* jury/*NN* commented/*VBD* on/*IN*
a/*DT* number/*NN* of/*IN* other/*JJ* topics/*NNS* ./.

Lematyzer

The/DT	man/NN	took/VB	the/DT	book/NN	./.
↓	↓	↓	↓	↓	↓
the	man	take	the	book	.

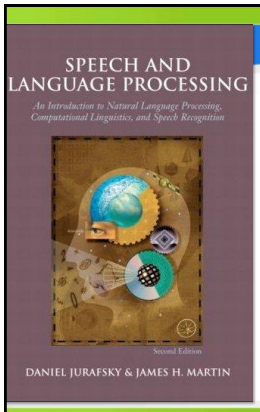
Drzewo rozbioru (*parse tree*)



Podsumowanie

- **Gramatyka formalna**
- **Wywód zdania**
- **Język formalny**
- **Gramatyki bezkontekstowe**
- **Notacja EBNF**

Literatura



Daniel Jurafsky, James Martin: Speech and Language Processing, Prentice-Hall, 2008.