


Introduction to Informatics

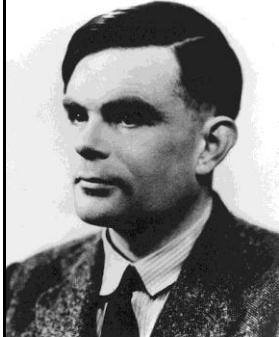
**Jerzy Nawrocki**  
Faculty of Computing and Inf. Sci.  
Poznan University of Technology  
jerzy.nawrocki@put.poznan.pl

**Artificial Intelligence and Natural Language Processing**



Introduction to Informatics

**Turing's Test (1950)**



**Alan Turing**


Artificial Intelligence ... (2)

A diagram illustrating Turing's Test. It shows a box labeled 'Computer' (A) and a box labeled 'Man' (B) at the top. Lines from both boxes converge on a box labeled 'Player' at the bottom. A horizontal line is drawn between the 'Computer' and 'Man' boxes.

Introduction to Informatics

**The ELIZA Program (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**

Artificial Intelligence ... (3)

Introduction to Informatics

**Aim**



- Introduction to natural language processing
- The concept of grammar

Artificial Intelligence ... (4)

Introduction to Informatics

**Agenda**




- Introduction
- Parts of Speech
- Grammar and derivation
- Formal language
- Context-free grammars
- Backus-Naur Form

Artificial Intelligence ... (5)

Introduction to Informatics

**Dionysios Thrax (170 – 90 BC)**



- 8 Parts of speech, POS:
  - Noun
  - Verb
  - Pronoun
  - Preposition
  - Adverb
  - Conjunction
  - Participle
  - Interjection
- Adjective?

<http://www.uni-koeln.de/phil-fak/ifa/NRWakademie/papyrologie/PK0eln/PK5128v.jpg>

Artificial Intelligence ... (6)

Introduction to Informatics

### Contemporary classification – English

- Noun
- Verb
- Pronoun
- Preposition
- Adverb
- Conjunction
- Participle
- Interjection

Parts of speech ↔ tags

Penn Treebank: 45 tags  
 Tags C5: 61 tags  
 Brown corpus: 87 tags  
 Tags C7: 146 tags

Artificial Intelligence ... (7)

Introduction to Informatics

### Penn Treebank Tags

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*

Artificial Intelligence ... (8)

Introduction to Informatics

### Tagging parts of speech

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 ./.

Artificial Intelligence ... (9)

Introduction to Informatics

### Tagging parts of speech

Seq. of words → Tagging parts-of-speech → Tagged words

↑ Classification of parts of speech

Artificial Intelligence ... (10)

Introduction to Informatics

### Tagging parts of speech – Main problem

Book that flight.

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

Other possibilities:  
 Book/NN *Buy me that book.*  
 that/CC *I thought that you knew.*

## Ambiguity

Artificial Intelligence ... (11)

Introduction to Informatics

### Removing ambiguity

Book that flight.

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

(A book) that/DT flight. } Nonsens  
 (A book), that/CC flight. }

Rule 1:  
 If there is a conflict between a verb (VB, VBD, ..) and another part of speech, and in the sentence there is no other verb, then accept that word as a verb.

Artificial Intelligence ... (12)

Introduction to Informatics  
Removing ambiguity

Book that flight.

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

(Make a reservation for) that/DT flight.  
~~(Make a reservation for), that/CC flight.~~

**Rule 2:**  
If there is a conflict between /CC and another part of speech and in the sentence there is only one verb, then the option /CC should be rejected.

Artificial Intelligence ... (13)

Introduction to Informatics  
Removing ambiguity – Rule-based approach

```

    graph LR
        A[Seq. of words] --> B[Initial part-of-speech tagging]
        C[POS classification] --> B
        B --> D[Words with many tags]
        D --> E[Reducer]
        F[A set of rules] --> E
        E --> G[Words with single tags]
    
```

The ENGTWOL system: about 1100 rules

Artificial Intelligence ... (14)

Introduction to Informatics  
Removing ambiguity

- Rule-based approach
- Stochastic approach (hidden Markov Models – HMM; frequency-based approach)

Artificial Intelligence ... (15)

Introduction to Informatics  
Agenda

- Introduction
- Parts of Speech
- Grammar and derivation
- Formal language
- Context-free grammars
- Backus-Naur Form

Artificial Intelligence ... (16)

Introduction to Informatics  
Part-of-speech tagging

```

    graph LR
        A[Words] --> B[Part-of-speech tagging]
        C[POS classification] --> B
        B --> D[Tagged words]
    
```

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

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

Artificial Intelligence ... (17)

Introduction to Informatics  
Part-of-speech tagging

```

    graph LR
        A[Words] --> B[Part-of-speech tagging]
        C[Parts of speech classification] --> B
        B --> D[Tagged words]
        D --> E[Further processing]
        F[?] --> E
    
```


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

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

Artificial Intelligence ... (18)

Introduction to Informatics

### Syntax of a sentence




Noam Chomsky

$S \rightarrow NP VP$      *We + are smart*

Artificial Intelligence ... (19)

Introduction to Informatics

### Noun phrase – NP



Noam Chomsky


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

POS: Noun (NN), ProperNoun (NNP), Det (NT), Pronoun  
Nonterminals (parts of sentence): NP, Nominal

Artificial Intelligence ... (20)

Introduction to Informatics

### Verb phrase – VP



Noam Chomsky


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

POS: VerbBase (VB), VerbPast (VBD)  
Nonterminals (parts of sentence): VP, Verb, NP

Artificial Intelligence ... (21)

Introduction to Informatics

### Grammar



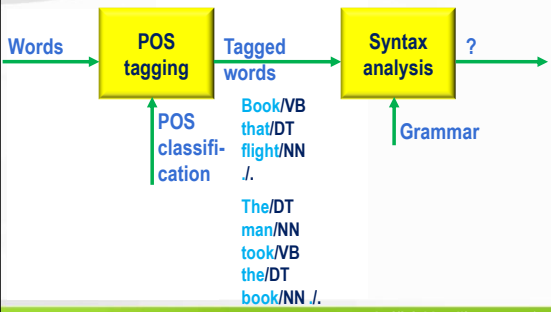
Noam Chomsky

- $S \rightarrow NP VP$
- NP → Pronoun
- | ProperNoun
- | Det Nominal
- Nominal → Noun
- | Noun Nominal
- VP → Verb
- | Verb NP
- Verb → VerbBase
- | VerbPast

Artificial Intelligence ... (22)

Introduction to Informatics

### POS tagging



Words → POS tagging → Tagged words → Syntax analysis → ?

POS classification

- Book/VB
- that/DT
- flight/NN
- .
- The/DT
- man/NN
- took/VB
- the/DT
- book/NN
- .

Grammar

Artificial Intelligence ... (23)

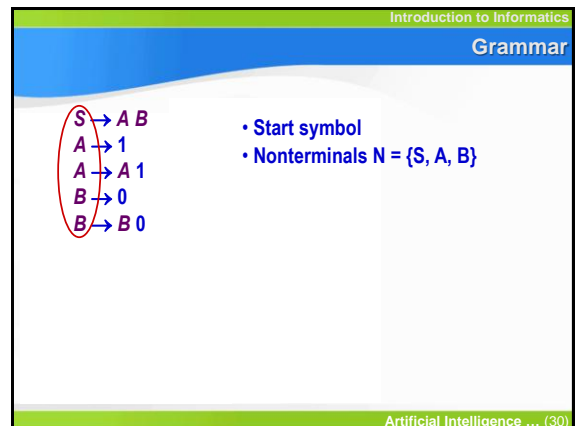
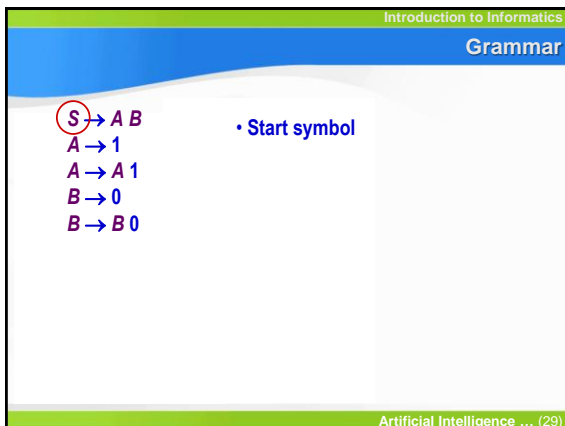
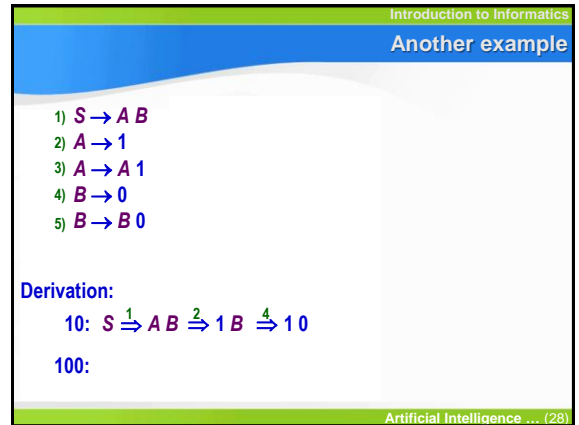
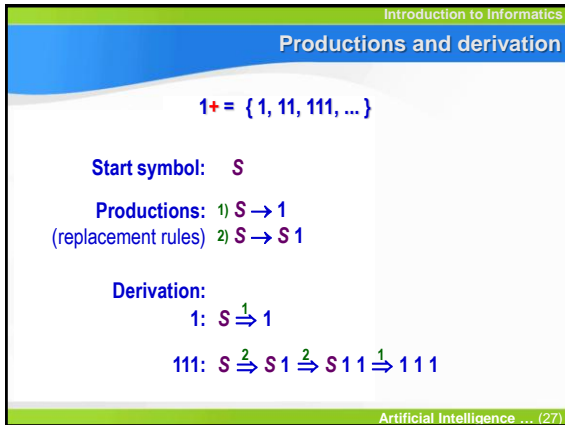
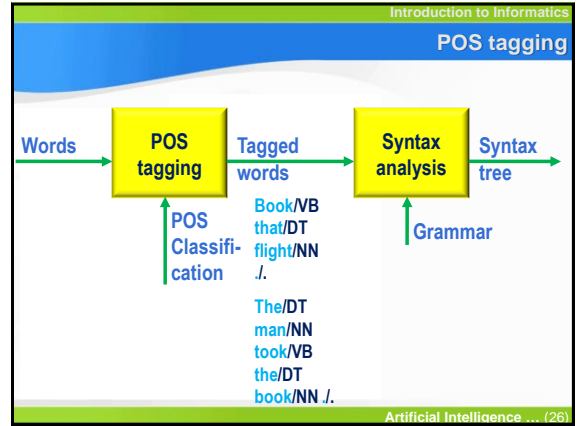
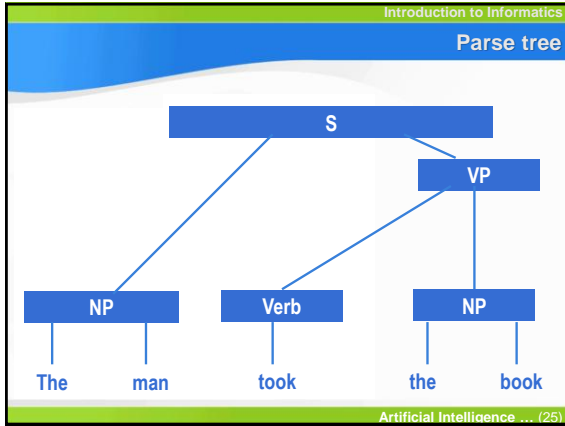
Introduction to Informatics

### Grammar

- $S \rightarrow NP VP$
- 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     ./.

Artificial Intelligence ... (24)



Introduction to Informatics  
Grammar

$S \rightarrow AB$   
 $A \rightarrow 1$   
 $A \rightarrow A1$   
 $B \rightarrow 0$   
 $B \rightarrow B0$

- Start symbol
- Nonterminals  $N = \{S, A, B\}$
- Terminals  $T = \{0, 1\}$

Artificial Intelligence ... (31)


Introduction to Informatics  
Grammars

$S \rightarrow AB$   
 $A \rightarrow 1$   
 $A \rightarrow A1$   
 $B \rightarrow 0$   
 $B \rightarrow B0$

- Start symbol
- Nonterminals  $N = \{S, A, B\}$
- Terminals  $T = \{0, 1\}$
- Productions

Artificial Intelligence ... (32)

Introduction to Informatics  
Agenda



- Introduction
- Parts of Speech
- Grammar and derivation
- Formal language
- Context-free grammars
- Backus-Naur Form

Artificial Intelligence ... (33)

Introduction to Informatics  
Closure of the derivation relation

- 1)  $S \rightarrow AB$
- 2)  $A \rightarrow 1$
- 3)  $A \rightarrow A1$
- 4)  $B \rightarrow 0$
- 5)  $B \rightarrow B0$

Derivation:

$S \xRightarrow{1} AB \xRightarrow{2} 1B \xRightarrow{4} 10$   
 $S \xrightarrow{+} 10$  From  $S$  one can derive  $10$  applying 1 or more productions

Artificial Intelligence ... (34)

Introduction to Informatics  
A set of strings over an alphabet

$S \rightarrow AB$   
 $A \rightarrow 1$   
 $A \rightarrow A1$   
 $B \rightarrow 0$   
 $B \rightarrow B0$

Alphabet = A set of terminal symbols  
 $T = \{0, 1\}$

Set of strings over an alphabet  $T^*$ :  
 Set of all finite strings built from elements belonging to  $T$ .

If  $T = \{0, 1\}$  then  $T^* = \{\epsilon, 0, 1, 00, 01, 10, 11, 000, \dots\}$   
 If  $T = \{a, b, c\}$  then  $T^* = \{\epsilon, a, b, c, aa, ab, ac, ba, bb, bc, \dots\}$

Artificial Intelligence ... (35)

Introduction to Informatics  
Formal language

Grammar  $G = \langle S, N, T, P \rangle$

$S$  – Start symbol  
 $N$  – Nonterminals  
 $T$  – Terminals  
 $P$  – Set of productions

Formal language  $L$  defined by a grammar  $G$ :

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

Artificial Intelligence ... (36)



Introduction to Informatics  
Formal language

- 1)  $S \rightarrow AB$
- 2)  $A \rightarrow 1$
- 3)  $A \rightarrow A1$
- 4)  $B \rightarrow 0$
- 5)  $B \rightarrow B0$

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


---


$$S \xrightarrow{1} AB \xrightarrow{2} 1B$$

$$S \xrightarrow{+} 1B$$

Does  $1B$  belong to  $L(G)$  ?

---


$$11 \in T^*$$

Does  $11$  belong to  $L(G)$  ?

Artificial Intelligence ... (37)

Introduction to Informatics  
Equivalence of grammars


Grammars  $G1$  i  $G2$  are equivalent iff:

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

G1	G2	G3
$S \rightarrow AB$	$S \rightarrow S0$	$S \rightarrow 1S$
$A \rightarrow 1$	$S \rightarrow A0$	$S \rightarrow 1A$
$A \rightarrow A1$	$A \rightarrow 1$	$A \rightarrow 0$
$B \rightarrow 0$	$A \rightarrow A1$	$A \rightarrow 0A$
$B \rightarrow B0$		

Artificial Intelligence ... (38)


Introduction to Informatics  
Agenda



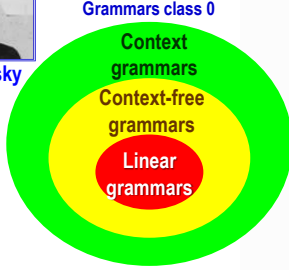
- Introduction
- Parts of Speech
- Grammar and derivation
- Formal language
- Context-free grammars
- Backus-Naur Form

Artificial Intelligence ... (39)

Introduction to Informatics  
Chomsky's hierarchy of grammars



Noam Chomsky




Grammars class 0

- Context grammars
- Context-free grammars
- Linear grammars

Artificial Intelligence ... (40)

Introduction to Informatics  
Chomsky's hierarchy of grammars



Linear grammars

Artificial Intelligence ... (41)

Introduction to Informatics  
Linear grammars

$a+ b+$


<ol style="list-style-type: none"> <li>1. <math>S \rightarrow S b</math></li> <li>2. <math>S \rightarrow A b</math></li> <li>3. <math>A \rightarrow a</math></li> <li>4. <math>A \rightarrow A a</math></li> </ol> <p>Left-recursive</p>		<ol style="list-style-type: none"> <li>1. <math>S \rightarrow a S</math></li> <li>2. <math>S \rightarrow a B</math></li> <li>3. <math>B \rightarrow b B</math></li> <li>4. <math>B \rightarrow b</math></li> </ol> <p>Right-recursive</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Theorem.**  
For every regular expression there is a left-recursive (right-recursive) grammar describing the same language.


Artificial Intelligence ... (42)

Introduction to Informatics

### Chomsky's hierarchy of grammars



Noam Chomsky



Context-free grammars  
Linear grammars

Artificial Intelligence ... (43)

Introduction to Informatics

### Context-free grammar

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

One nonterminal

Artificial Intelligence ... (44)

Introduction to Informatics

### Context-free grammar


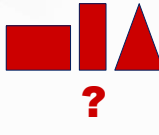

1.  $W \rightarrow S$
2.  $W \rightarrow W + S$
3.  $S \rightarrow C$
4.  $S \rightarrow S * C$
5.  $C \rightarrow L$
6.  $C \rightarrow ( W )$
7.  $L \rightarrow 1$
8.  $L \rightarrow 2$
9.  $L \rightarrow 3$

One nonterminal

Artificial Intelligence ... (45)

Introduction to Informatics

### Context-free grammar are more expressive

Context-free grammars  
Linear grammars

Artificial Intelligence ... (46)

Introduction to Informatics

### Context-free grammar are more expressive

Język  $0^n 1^n$ , gdzie  $n \geq 1$ .

0011 OK.

0001 Error

$S \rightarrow 0 S 1$

$S \rightarrow 0 1$

Język  $0^n 1^k$ , gdzie  $n, k \geq 1$ .

0001 OK.


1000 Error

$S \rightarrow 0 S$

$S \rightarrow 0 J$

$J \rightarrow 1 J$

$J \rightarrow 1$



Context-free grammars  
Linear grammars


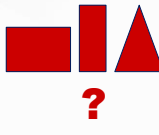

Artificial Intelligence ... (47)

Introduction to Informatics

### Context-free grammar are more expressive

Język  $0^n 1^n$ , gdzie  $n \geq 1$ .

Język  $0^n 1^k$ , gdzie  $n, k \geq 1$ .

Context-free grammars  
Linear grammars

Artificial Intelligence ... (48)



Introduction to Informatics

### Chomsky's hierarchy of grammars

Grammars class 0

- Context grammars
- Context-free grammars
- Linear grammars

Noam Chomsky

Artificial Intelligence ... (49)

Introduction to Informatics

### Context grammars

- $S \rightarrow aXY$
- $S \rightarrow aSXY$
- $aX \rightarrow ab$
- $bX \rightarrow bb$
- $cX \rightarrow cc$
- $bY \rightarrow bc$
- $cY \rightarrow cc$

Artificial Intelligence ... (50)

Introduction to Informatics

### Agenda

- Introduction
- Parts of Speech
- Grammar and derivation
- Formal language
- Context-free grammars
- Backus-Naur Form

Artificial Intelligence ... (51)

Introduction to Informatics

### Extended Backus-Naur Form

Productions + Regular expressions

$\langle C \rangle ::= '0' | '1' | '2' | '3' | '4' | '5' | '6' | '7' | '8' | '9'$

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

$\langle S \rangle ::= \langle C \rangle^* \langle C \rangle$

$\langle S \rangle ::= (\langle L \rangle^{**})^* \langle L \rangle$

$\langle W \rangle ::= (\langle S \rangle '+' )^* \langle S \rangle$

John Backus

Artificial Intelligence ... (52)

Introduction to Informatics

### From EBNF to grammars

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

$J \rightarrow B$

$J \rightarrow AJ$

Artificial Intelligence ... (53)

Introduction to Informatics

### From EBNF to grammars

$\langle C \rangle ::= '0' | '1' | '2' | '3' | '4' | '5' | '6' | '7' | '8' | '9'$

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

$\langle S \rangle ::= (\langle L \rangle^{**})^* \langle L \rangle$

$\langle W \rangle ::= (\langle S \rangle '+' )^* \langle S \rangle$

$C \rightarrow '0'$

$C \rightarrow '1'$

$C \rightarrow '2'$

...

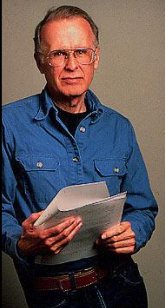
$C \rightarrow '9'$

John Backus

Artificial Intelligence ... (54)

Introduction to Informatics

### From EBNF to grammars



```

<C> ::= '0' | '1' | '2' | '3' | '4' | '5' | '6' | '7' |
      '8' | '9'
<L> ::= <C>* <C>
<S> ::= (<L> {8*})* <L>
<W> ::= (<S> '+' )* <S>

<J> ::= <A>* <B>
J → B
J → A J

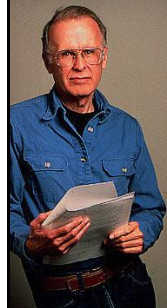
L → C
L → C L
    
```

John Backus

Artificial Intelligence ... (55)

Introduction to Informatics

### From EBNF to grammars



```

<C> ::= '0' | '1' | '2' | '3' | '4' | '5' | '6' | '7' |
      '8' | '9'
<L> ::= <C>* <C>
<S> ::= (<L> {8*})* <L>
<W> ::= (<S> '+' )* <S>

<J> ::= <A>* <B>
J → B
J → A J

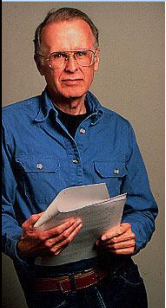
S → L
S → L {8*} S
    
```

John Backus

Artificial Intelligence ... (56)

Introduction to Informatics

### From EBNF to grammars



```

<C> ::= '0' | '1' | '2' | '3' | '4' | '5' | '6' | '7' |
      '8' | '9'
<L> ::= <C>* <C>
<S> ::= (<L> {8*})* <L>
<W> ::= (<S> '+' )* <S>


<J> ::= <A>* <B>
J → B
J → A J

W → S
W → S '+' W
    
```

John Backus

Artificial Intelligence ... (57)

Introduction to Informatics




### Summary

Artificial Intelligence ... (58)

Introduction to Informatics

### Aim

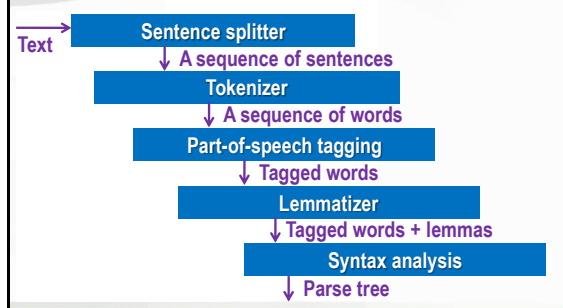


- Introduction to natural language processing
- The concept of grammar

Artificial Intelligence ... (59)

Introduction to Informatics

### An exemplary architecture of an NLP tool



```

Text → Sentence splitter
      ↓ A sequence of sentences
      Tokenizer
      ↓ A sequence of words
      Part-of-speech tagging
      ↓ Tagged words
      Lemmatizer
      ↓ Tagged words + lemmas
      Syntax analysis
      ↓ Parse tree
    
```

Artificial Intelligence ... (60)

Introduction to Informatics  
Part-of-speech tagging

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 ./.

Artificial Intelligence ... (61)

Introduction to Informatics  
Lemmatizer

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

the man take the book .

Artificial Intelligence ... (62)

Introduction to Informatics  
Parse tree

```

    graph TD
      S[S] --- NP1[NP]
      S --- VP[VP]
      NP1 --- The[The]
      NP1 --- man[man]
      VP --- took[took]
      VP --- NP2[NP]
      NP2 --- the[the]
      NP2 --- book[book]
  
```

Artificial Intelligence ... (63)

Introduction to Informatics  
Summary

- Formal grammar
- Derivation of a sentence
- Formal language
- Context-free grammars
- EBNF notation

Artificial Intelligence ... (64)

Introduction to Informatics  
Bibliography

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

Artificial Intelligence ... (65)

Introduction to Informatics

**Thank you for your attention!**

Artificial Intelligence ... (66)