

ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

Комуникацијата меѓу агентите

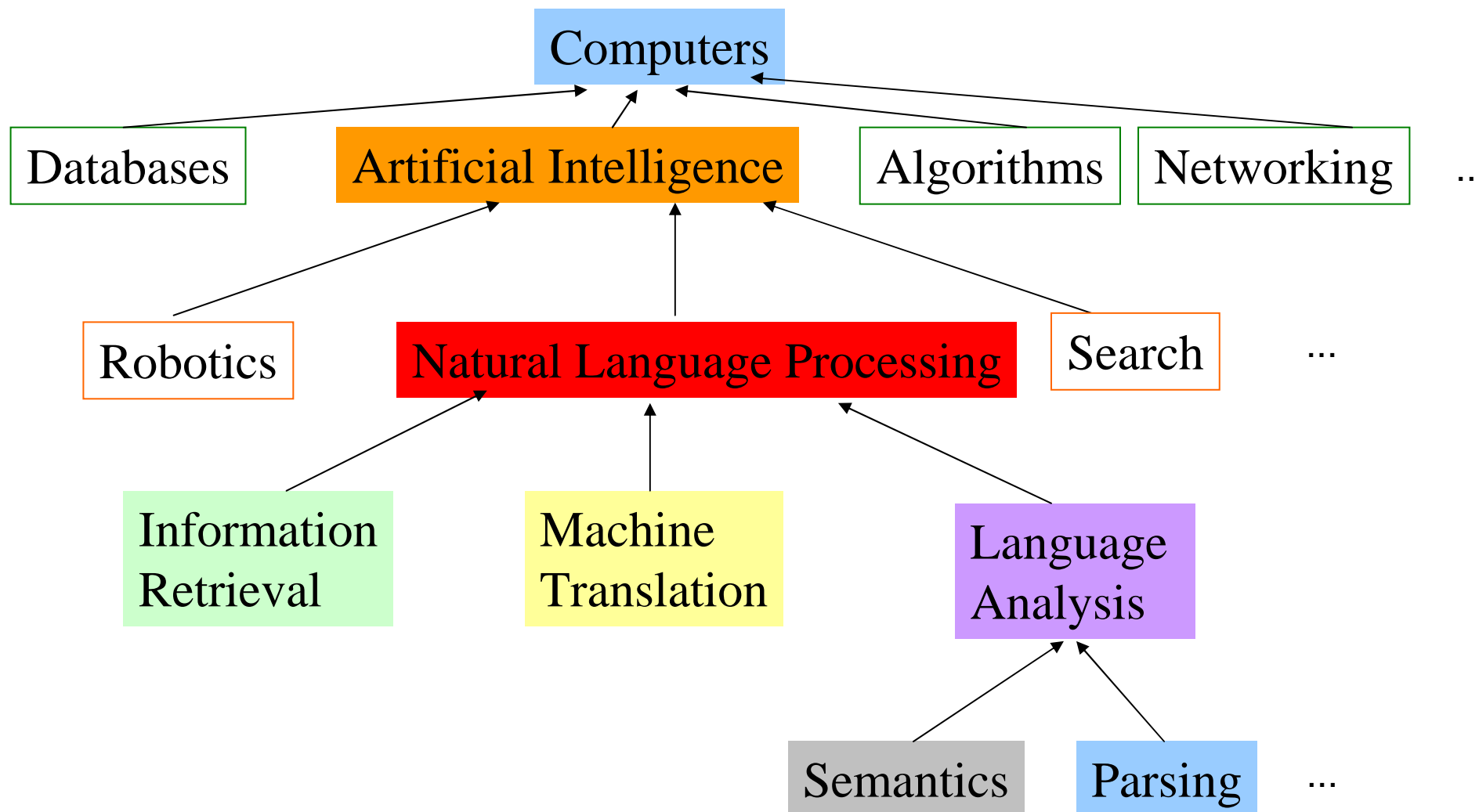
Комуникација

- Дефиницијата од Википедија:
Комуникацијата претставува процес на пренос на информациите од испраќачот кон примателот со помош на таков медиум преку кој информацијата која се препраќа ја разбираат и испраќачот и примателот.
- Притоа, ако комуникацијата се одвива меѓу луѓето, тогаш испраќачот е говорник, медиумот се акустичните бранови, а примателот е слушателот.

Обработка на природни јазици (Natural Language Processing)

- Првенствено се проучуваат сите нешта што се неопходни за да се овозможи заемодејство (интеракција) со компјутерите на многу попрроден начин (со користење на природен јазик)
- Но, исто така, ваквите компјутерски обработки на природните јазици, даваат идеи за можните начини како луѓето ги обработуваат јазиците

Каде се вклопува во поделбата на компјутерските науки?



Актуелни примени на обработката на природни јазици

- Одговарање на прашања
(Question answering)
- Конверзациски агенти
(Conversational agents)
- Прераскажување (резимирање)
(Summarization)
- Анализа на оговарања и расположение
(Buzz analysis, Sentiment analysis)
- Машинско преведување
(Machine translation)

Трите основни семиотски правила

- Семиотиката е наука која се занимава со проучување на знаковните процеси и со тоа како се создава и сфаќа значењето.
- Трите основни правила на семиотиката се:
 1. синтаксата
 2. прагматиката и
 3. семантиката

Синтакса

- Синтаксата е наука која се занимава со наоѓањето на правилата според кои се конструираат речениците во еден јазик.
- На пример, правилото според кое е конструирана реченицата “Времето е многу топло.” е: именка, глагол, прилог, придавка.
- А правилото на реченицата “Топлото време им пречи на луѓето.” е: придавка, именка, кратка заменска форма, глагол, предлог, именка.

Прагматика

- Прагматиката е филозофски правец кој се занимава со тоа на каков начин се користат зборовите при комуникацијата.
- На пример, реченицата “Моливот ја преврте кравата” е граматички сосема исправна, но не носи порака, бидејќи навидум нема смисол. Но, ако на оваа реченица и претходи реченицата “Весна го фрли моливот кон гумената крава.”, тогаш реченицата станува прифатлива.

Семантика

- Семантиката е наука која се занимава со значењето на зборовите (лексичка семантика), комбинациите зборови (фразеологија) и како фразите се комбинираат за да може да се одреди значењето на реченицата.
- На пример, дали го знаете значењето на реченица која започнува со: “Петар јаде буреќ со ...”, а на местото на празните места има само една именка?

Петар јаде бурека со ...

- месо.
- прсти.
- Марко.
- јогурт.
- неконгруентен атрибут
- инструмент
- објект (субјект)
- објект (објект)
- Секоја од четирите реченици има еднаква синтакса, но семантиката им е сосема различна.

Претставување на синтаксата

BNF = Backus-Naur form

- Реченицата се разбива на помали целини (нетерминални симболи).
- Тие постепено се трансформираат кон поелементарни нетерминални симболи кои ги претставуваат зборовните групи (PoS = Part of Speech).
- На крајот, овие нетерминални симболи се трансформираат во терминални симболи кои ги претставуваат зборовите во реченицата.

Силното сонце го стопи снегот.

Подготовка за BNF:

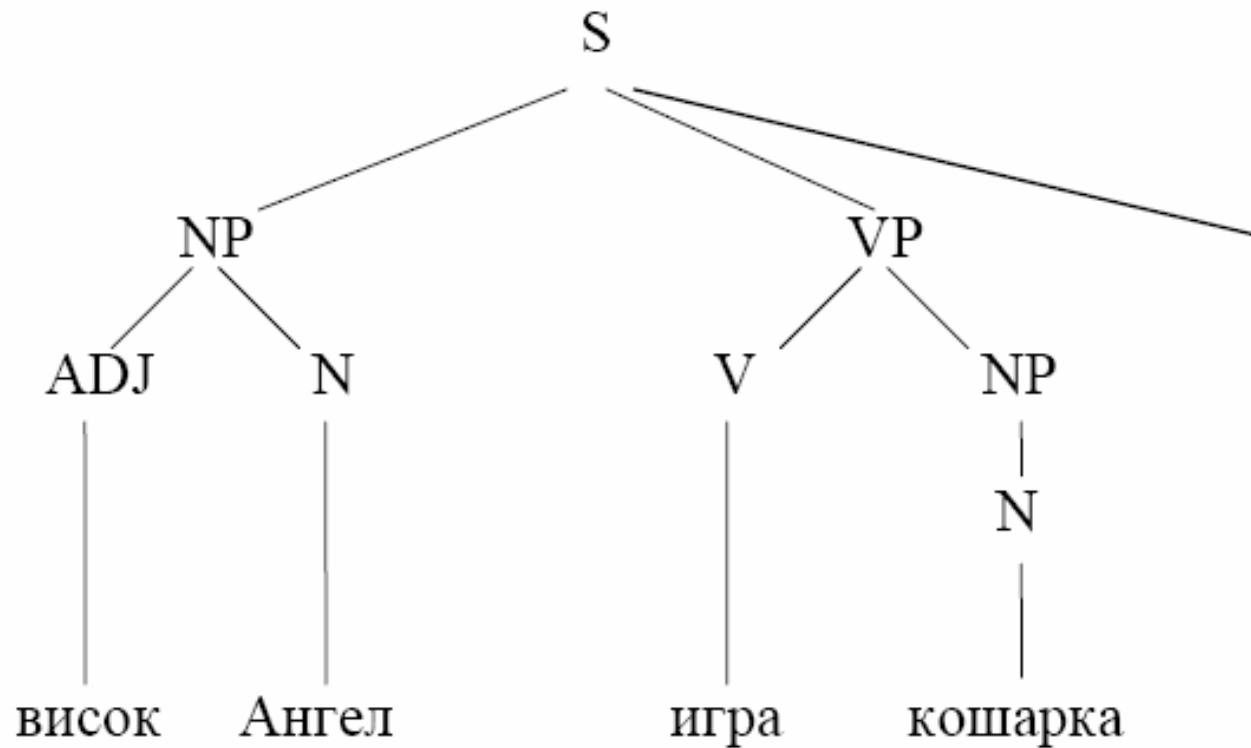
- $S = \text{подмет (NP)} + \{\text{прирок} + \text{предмет}\} (VP)$
- NP = силното сонце
- VP = го стопи снегот
- силното – придавка Adj
- сонце – именка N
- го – кратка заменска форма Pron
- стопи – глагол V
- снегот – именка N

Високиот Ангел игра кошарка.

Подготовка за BNF:

- $S = \text{подмет (NP)} + \{\text{прирок} + \text{предмет}\} (VP)$
- NP = високиот Ангел
- VP = игра кошарка
- високиот – придавка Adj
- Ангел – именка N
- игра – глагол V
- кошарка – именка N

Графички приказ: длабока структура (deep structure)



Некои поважни ознаки во BNF

- NP – noun phrase (именска фраза)
- VP – verb phrase (глаголска фраза)
- N – noun (именка)
- V – verb (глагол)
- Adj – adjective (придавка)
- Adv – adverb (прилог)
- Prep – Preposition (предлог)
- Conj – conjunction (сврзник)
- PP – personal pronoun (лична заменка)
- Det – determiner (член)

Контекстно-слободни граматики

- **N** множество нетерминални симболи (или променливи)
- Σ множество терминални симболи (различно од N)
- **R** множество од продукции или правила од обликот $A \rightarrow \alpha$, каде A е нетерминален симбол, а α е стринг од симболи од $(\Sigma \cup N)^*$
- **S** предодреден нетерминален симбол наречен почетен симбол

Контекстно-слободни граматики

Граматика

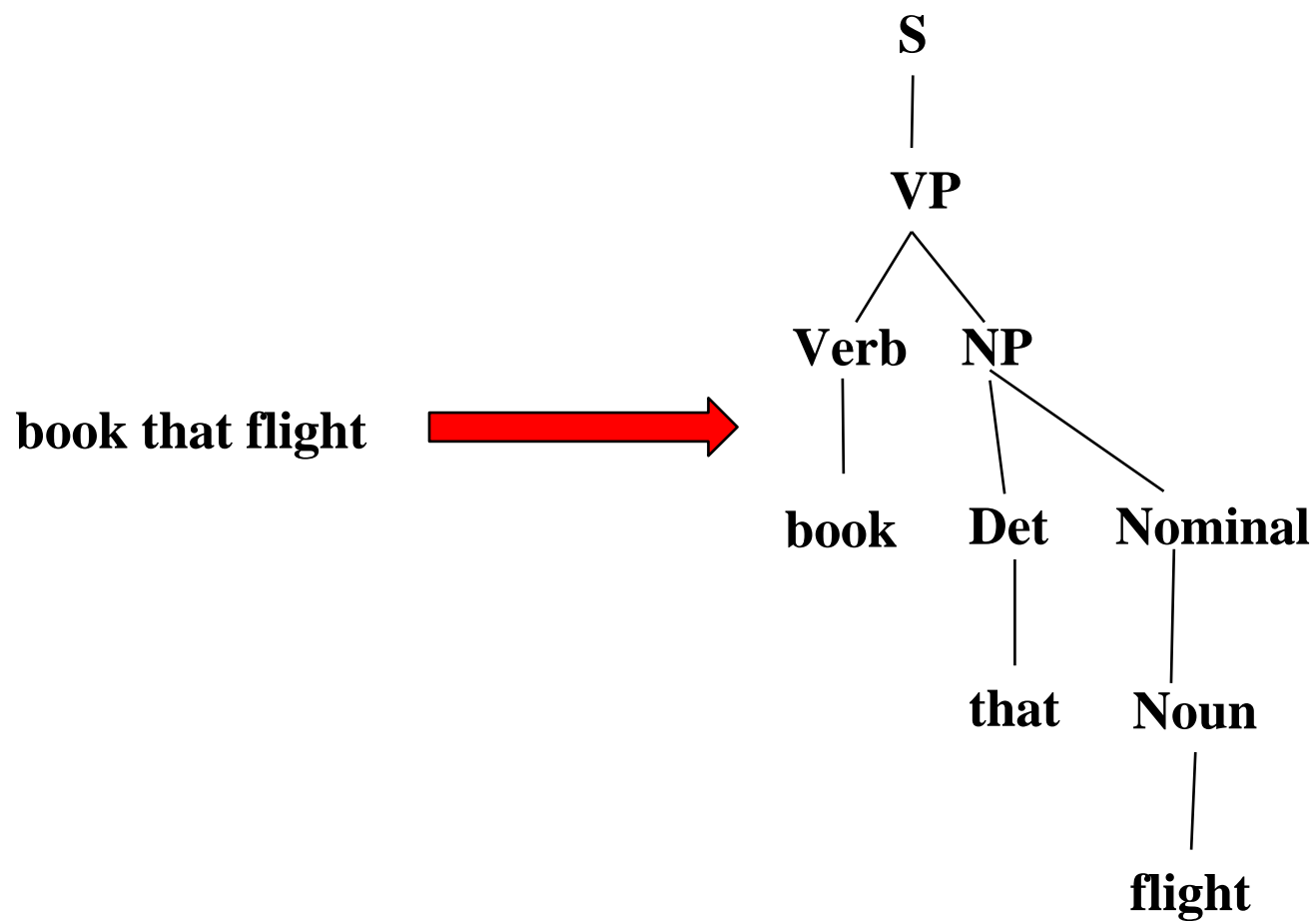
S → **NP VP**
S → **Aux NP VP**
S → **VP**
NP → **Pronoun**
NP → **Proper-Noun**
NP → **Det Nominal**
Nominal → **Noun**
Nominal → **Nominal Noun**
Nominal → **Nominal PP**
VP → **Verb**
VP → **Verb NP**
VP → **VP PP**
PP → **Prep NP**

Лексикон

Det → **the | a | that | this**
Noun → **book | flight | meal | money**
Verb → **book | include | prefer**
Pronoun → **I | he | she | me**
Proper-Noun → **Houston | NWA**
Aux → **does**
Prep → **from | to | on | near | through**

Пример граматика и лексикон
за корпусот ATIS за контрола
на авионски летови

Пример



Парсирање

- Одредувањето на синтаксичката исправност на реченицата се вика парсирање.
- Тоа значи дека се одредува дали редоследот на зборовните групи е точен или не.
- Парсирањето може да биде надолно (Top down) или нагорно (Bottom up)
- При надолното парсирање се тргнува од S кон збороформите
- При нагорното парсирање се тргнува од збороформите кон S

Пример

Дадена е следнава регуларна граматика:

$\langle S \rangle ::= \langle NP \rangle \langle VP \rangle$

$\langle NP \rangle ::= \langle ADJ \rangle \langle NP \rangle \mid \langle N \rangle$

$\langle VP \rangle ::= \langle ADV \rangle \langle VP \rangle \mid \langle V \rangle \langle NP \rangle \mid \langle V \rangle$

$\langle N \rangle ::= \text{Mare} \mid \text{Kiro} \mid \text{Toni} \mid \text{pesna} \mid \text{slatko}$

$\langle V \rangle ::= \text{pee} \mid \text{jade}$

$\langle ADJ \rangle ::= \text{ubavo} \mid \text{golemo} \mid \text{slatko} \mid \text{vkusno}$

$\langle ADV \rangle ::= \text{ubavo} \mid \text{vkusno} \mid \text{brzo}$

Дали со неа се парсира реченицата
“Маре јаде убаво вкусно слатко слатко”?

<S> => <NP> <VP>

=> <N> <VP>

=> Mare <VP>

=> Mare <V> <NP>

=> Mare jade <NP>

=> Mare jade <ADJ> <NP>

=> Mare jade ubavo <NP>

=> Mare jade ubavo <ADJ> <NP>

=> Mare jade ubavo вкусно <ADJ> <NP>

=> Mare jade ubavo вкусно слатко <NP>

=> Mare jade ubavo вкусно слатко <N>

=> Mare jade ubavo вкусно слатко слатко

Нагорно парсирање на реченицата “Маре јаде убаво вкусно слатко слатко”?

Mare \leq $\langle N \rangle$

jade \leq $\langle V \rangle$

ubavo \leq $\langle ADV \rangle$ | $\langle ADJ \rangle$

vkusno \leq $\langle ADV \rangle$ | $\langle ADJ \rangle$

slatko \leq $\langle N \rangle$ | $\langle ADJ \rangle$

$\langle N \rangle \leq \langle NP \rangle$

оттука:

slatko $\leq \langle NP \rangle$

$\langle V \rangle \langle ADV \rangle$, $\langle V \rangle \langle ADJ \rangle$

отпаѓаат

$\langle ADV \rangle \langle ADV \rangle$, $\langle ADV \rangle \langle ADJ \rangle$,
 $\langle ADJ \rangle \langle ADV \rangle$ и $\langle ADJ \rangle \langle ADJ \rangle$
исто така отпаѓаат

Како пар ADJ и NP го редуцираат
NP.

$\langle ADJ \rangle \langle NP \rangle \leq \langle NP \rangle$

Досега се извршени следниве
редукции:

Mare $\leq \langle NP \rangle$ jade \leq
 $\langle V \rangle$

slatko $\leq \langle NP \rangle$ slatko

slatko $\leq \langle NP \rangle$

$\langle ADV \rangle \langle NP \rangle$ отпаѓа

Од тоа што постои редукцијата:

$\langle ADJ \rangle \langle NP \rangle \leq \langle NP \rangle$

се генерира:

ubavo vkusno slatko slatko $\leq \langle NP \rangle$

Следуваат:

$\langle V \rangle \langle NP \rangle \leq \langle VP \rangle$

$\langle NP \rangle \langle VP \rangle \leq \langle S \rangle$

Задача:

Користејќи ја истата регуларна граматика, да се изврши нагорно и надолно парсирање на речениците:

1. Тони убаво пее убава песна.
2. Киро убаво слатко јаде големо вкусно Тони.
3. Убавата Маре убаво пее убава песна.
4. Маре јаде убаво.

Повеќезначност на толкувањето (1)

- Што значи фразата:
EXTRA WIDE VEHICLE ENTRANCE
- extra (wide (vehicle entrance))
Дополнителен широк влез за возила
- (extra wide) (vehicle entrance))
Исклучително широк влез за возила
- extra (wide vehicle) entrance
Дополнителен влез за широки возила

Повеќезначност на толкувањето (2)

- Што значи реченицата:

На полицајците им беше наредено да го прекинат пиењето во 12 часот.

- Тие пиеле па им било наредено да престанат да пијат.
- Им било наредено да влезат во крчмата во 12 часот и да направат фајронт.
- Во 12 часот им било наредено да направат фајронт.

Природните јазици и комуникацијата

- Трите основни правци на обработката на природните јазици (**Natural Language Processing, NLP**) се:
 1. Разбирањето / сфаќањето (**Understanding / Comprehension**)
 2. Генерирањето / создавањето (**Generation / Production**)
 3. Преведувањето (**Translation**)

Разбирање

- Препознавање на говорот
(Speech recognition)
- Синтаксичка анализа
(Syntactic analysis)
- Семантичка анализа
(Semantic analysis)
- Прагматична анализа
(Pragmatic analysis)

Генерирање и преведување

■ Генерирање

- ☐ Избор на содржини (Content selection)
- ☐ Синтаксичка реализација (Syntactic realization)
- ☐ Синтеза на говорот (Speech synthesis)

■ Преведување

- ☐ Разбирање (Understanding)
- ☐ Генерирање (Generation)

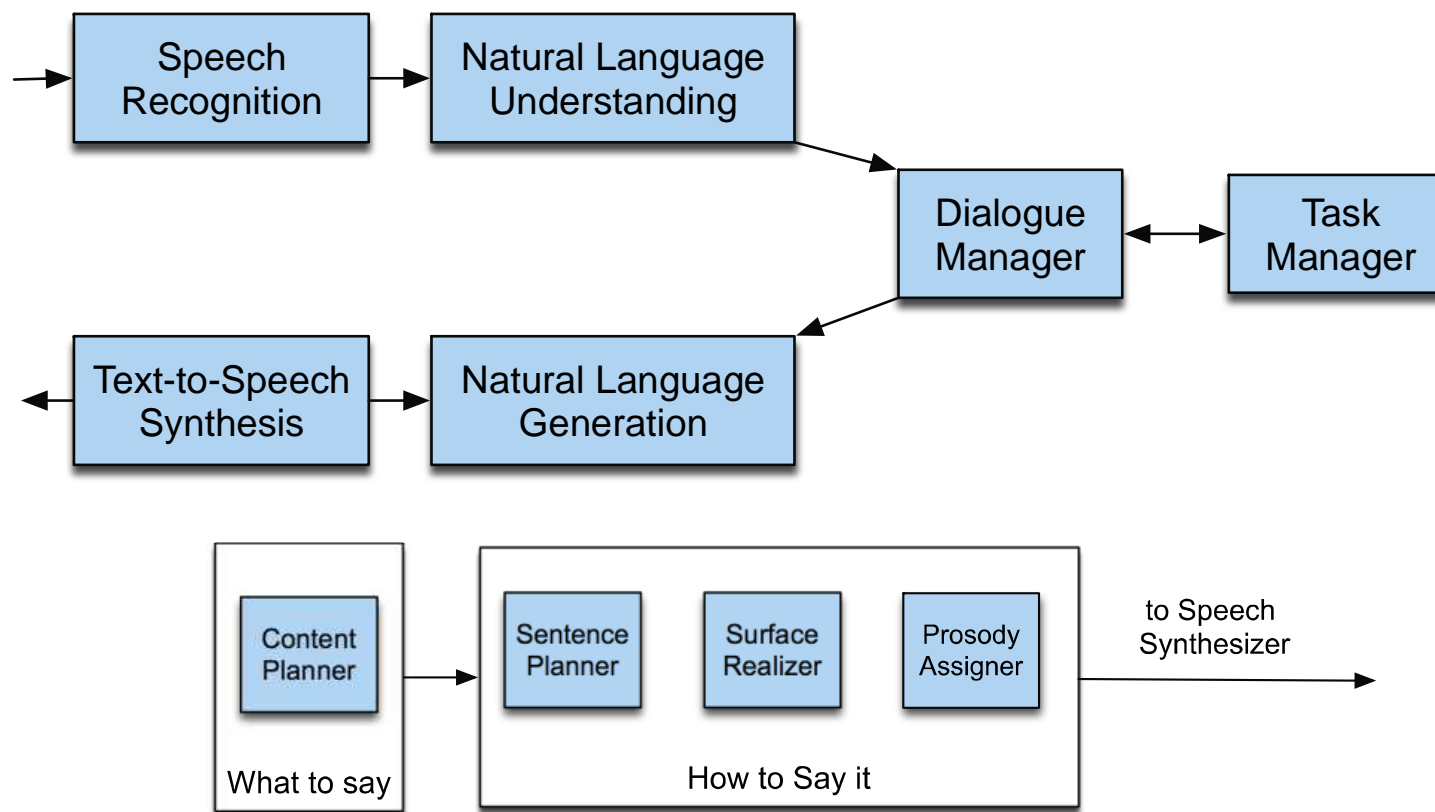
Разговорни агенти

- Што претставува еден разговорен дијалог систем?
 - Систем кој му овозможува на корисникот да ги изговара неговите прашања на природен јазик и на сличен начин да го добива одговорот од системот.
 - Говорните дијалог системи обезбедуваат интерфејс помеѓу корисникот и компјутерски-базираната апликација која дозволува говорна интеракција со апликацијата на релативно природен начин.



Разговорни агенти

■ Архитектура на еден разговорен агент



Повеќезначност и хомонимија

- Два збора со иста форма и различни значења се повеќезначни.

Пример: Петре има убава коса.

коса (алатка)

коса (дел од главата)

- Две или повеќе збороформи со иста форма се хомонимни

Пример: убаво дете убаво пее

убаво (придавка, опис на именката)

убаво (прилог, опис на глаголот)

Зборовна повеќезначност

- Разрешување на повеќезначноста
(Word Sense Disambiguation, WSD)
- Најсоодветен пристап: научи го
системот да врши добра класификација
- Алгоритми за учење:
 - ☐ веројатносно расудување (класичен Bayes)
 - ☐ логика од прв ред
 - ☐ невронски мрежи

Пристап

- Во речникот на секој јазик се дадени зборовните групи (PoS) на заглавните зборови
- Одредувањето на точната зборовна група во рамките на реченицата е проблем на класификација
- Контекстот во кој се очекува повеќезначноста треба да се претстави со помош на множество својства за нивно разрешување (на пр. редоследот на зборовните групи во реченицата, или претходниот контекст)
- Обучувањето на класификацијата се врши со помош на множество за обука

Одредување на зборовната група (Part Of Speech Tagging)

- Да се класифицира секоја збороформа независно, користејќи ги информациите од речникот и информациите за збороформите во околината.
- Означувањето повторно претставува класификација, значи, може да се научи
- Класификацијата може да биде од лево надесно (forward) или од десно налево (backward)
- Најуспешни алгоритми се т.н. Веројатносни последователни модели (Probabilistic sequence models)

Морфологија

- Синтаксата не е можна без морфологијата
- Со морфологијата се истражуваат збороформите и збороформирањето
- Морфологијата и синтаксата ги одредуваат морфосинтаксичките описи (Morpho-Syntactic Description, MSD) на збороформите во јазикот
- Пример за MSD-ата на Multext: највеселава Aqsfs-p

Зборовните групи во Multext

<http://nl.ijs.si/ME/Vault/V3/msd/html/>

Part-of-Speech	Code	Atts
Noun	N	10
Verb	V	15
Adjective	A	12
Pronoun	P	17
Determiner	D	10
Article	T	6
Adverb	R	6
Adposition	S	4
Conjunction	C	7
Numeral	M	12
Interjection	I	2
Residual	X	0
Abbreviation	Y	5
Particle	Q	3

MSD-ата за македонските именки

Атрибут	Вредност	Пример	Код
Тип	општа (common)	човек, жена, дете	c
	сопствена (proper)	Иван, Вера, Скопје	p
Род	машки (masculine)	човек, Иван	m
	женски (feminine)	жена, Вера	f
	среден (neuter)	дете, Скопје	n
Број	еднина (singular)	човек, жена, дете	s
	множина (plural)	луѓе, жени, деца	p
	избројана множина (count)	човека	t
Падеж	номинатив (nominative)	човек, Иван, жена, дете	n
	вокатив (vocative)	човеку, жено	v
	j.c.(l.s.) е-вокатив (e-vocative)	човече, Иване	w
	j.c.(l.s.) а-форма (oblique)	човека, Ивана, ешка	o
Одреденост	не (no)	човек, жена, дете	n
	да (yes)	човекот, жената, детето	y
	j.c.(l.s.) блиска (proximal)	човеков, женава, детево	p
	j.c.(l.s.) далечна (distal)	човекон, женана, детено	d

4 Person	first	1	x	x	x	x	x	x	x	x	x	x	x
	second	2	x	x	x	x	x	x	x	x	x	x	x
	third	3	x	x	x	x	x	x	x	x	x	x	x

5 Number	singular	s	x	x	x	x	x	x	x	x	x	x	x
	plural	p	x	x	x	x	x	x	x	x	x	x	x
	dual	d			x								x
	collective	l											x

6 Gender	masculine	m		x	x	x	x	x			x	x	x
	feminine	f		x	x	x	x	x			x	x	x
	neuter	n		x	x	x	x	x			x	x	x
* ***** *													
7 Voice	active	a			x	x	x	x	x		x	x	x
	passive	p			x	x	x	x	x		x	x	x

8 Negative	no	n			x	x		x	x		x	x	x
	yes	y			x	x		x	x		x	x	x

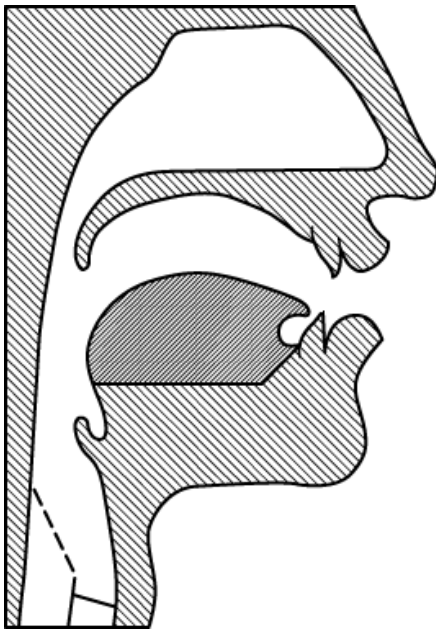
9 Definiteness	no	n					x				x		
	yes	y					x				x		
	short_art	s					x						
	full_art	f					x						

WordNet – семантички лексикон

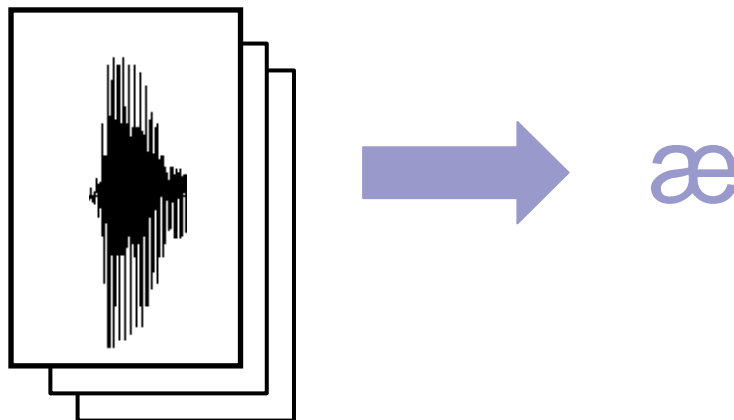
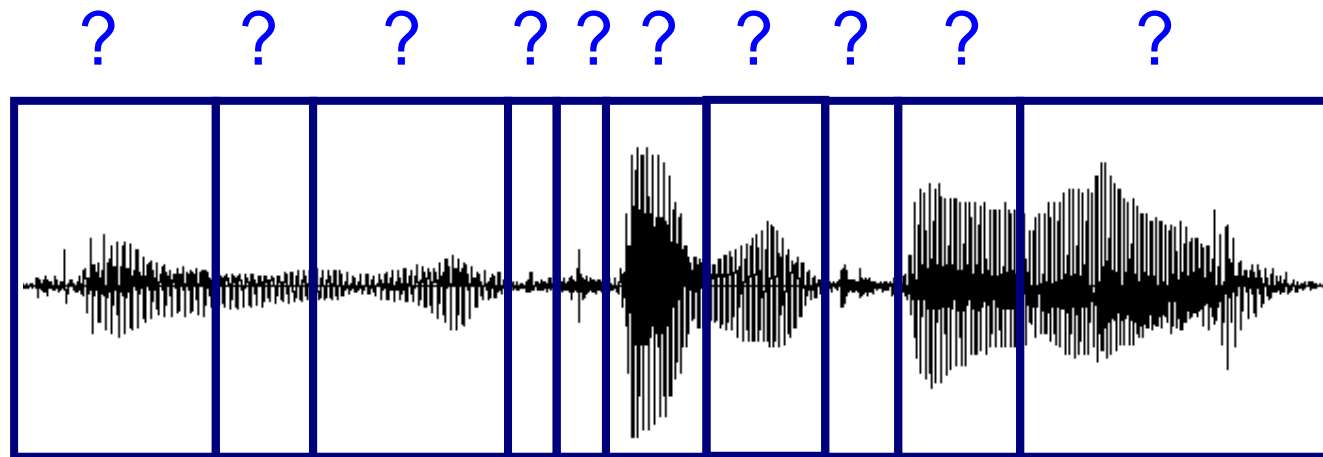
- WordNet е семантички лексикон на јазиците
- Целта му е да направи речник и тезаурус соодветен за интелигентна обработка со помош на компјутер.
- Групирање на зборовите во множества синоними (зборови со исто или слично значење), таканаречени synsets.
- Освен врска меѓу зборовите кои се синоними, за зборовите кои имаат антоними (зборови со спротивно значење) се воспоставуваат и односи кон антонимите.
- Во рамките на син-множествата, меѓу зборовите се дефинираат семантички односи.
- WordNet е организиран хиерархиски.

Акустично моделирање (Acoustic Modeling)

$$P(words)P(acoustics|words)$$

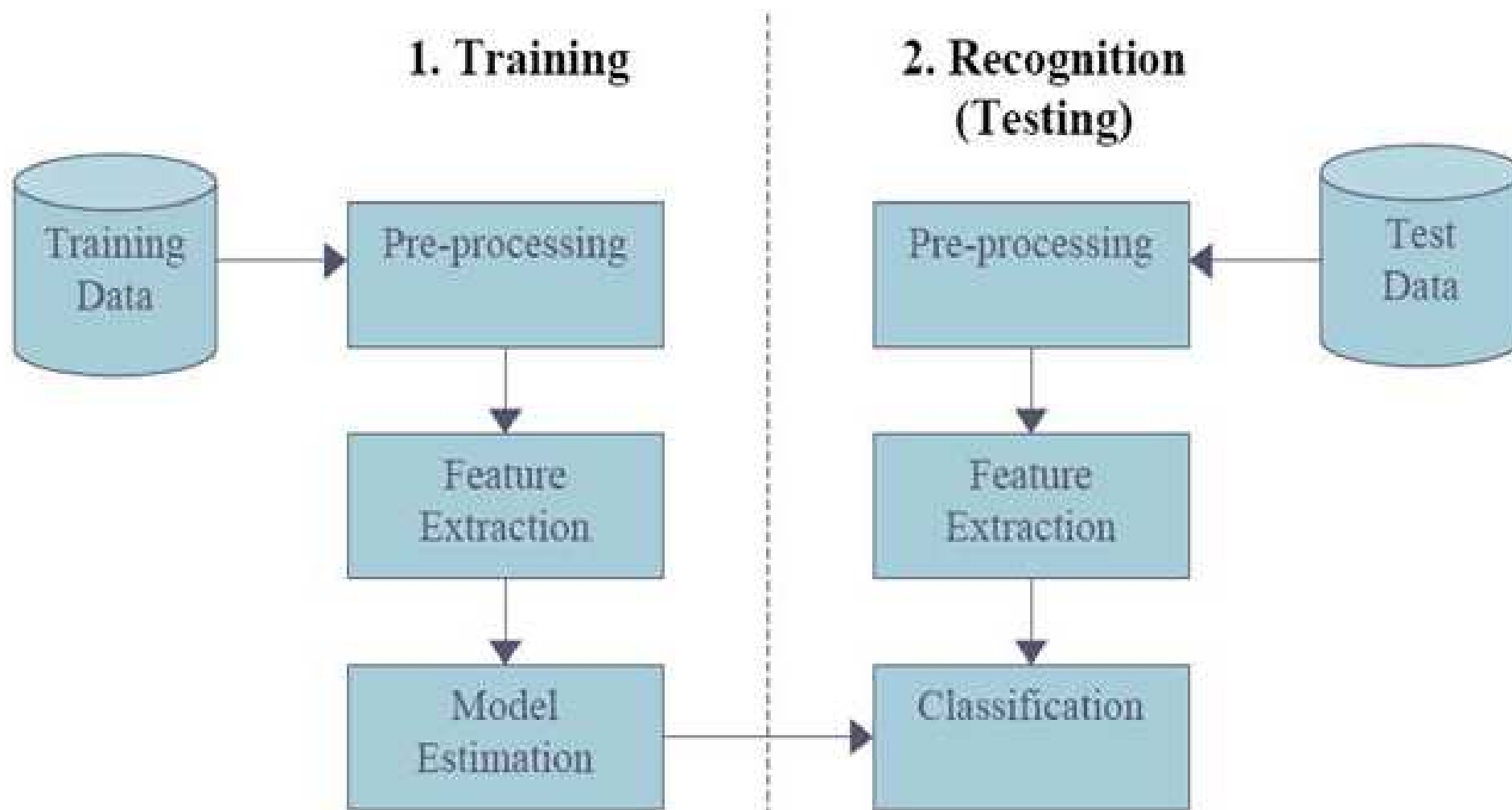


Класификација на фонеме



Нема јасна граница каде почнува дадена буква (поточно фонема), каде завршува.

Типичен систем за препознавање на говор



Делови на системот за препознавање на говор

- Предобработката (pre-processing) обично вклучува откривање на постоење на звук и отстранување на шум
- Извлекувањето на својства (feature extraction) вклучува разни трансформации од временски во фреквентен домен (со варијанти на Фуриеова трансформација), при што сигналот се дели на мали рамки од по 20 ms, поместени за 5 до 10 ms.
- Класификацијата и изградбата на т.н. модел обично се прави со невронски мрежи комбинирани со некаков систем за анализа на секвенци



Прашања?