

Roviny analýzy jazyka. Fonetika

Aleš Horák

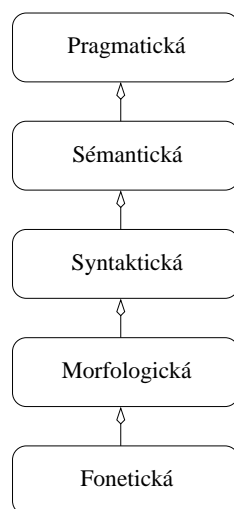
E-mail: hales@fi.muni.cz
http://nlp.fi.muni.cz/poc_lingv/

Obsah:

- ▶ Roviny analýzy jazyka
- ▶ Fonetika a fonologie

Roviny analýzy jazyka

znalosti struktury jazyka jsou propojeny **hierarchicky**



jazykové **roviny**:

- ▶ **fonetická**
- ▶ **morfologická**
- ▶ **syntaktická**
- ▶ **sémantická**
- ▶ **pragmatická**
- ▶ kontextová
- ▶ znalost základní ontologie
- ▶ jazykové metaznalosti

Struktura jazyka

Struktura jazyka zahrnuje informace o:

- ▶ co jsou **slova** (z jakých **znaků**, jaké slovní tvary a jejich složky)
- ▶ jak se slova (větné složky) kombinují do **vět**
- ▶ co slova označují, jaké jsou jejich **lexikální významy**
- ▶ jak se **význam věty** skládá z významů slov a slovních spojení (větných složek)

zpracování jazyka dále potřebuje:

- ▶ obecnou (encyklopedickou) **znalost světa** (ontologie)
- ▶ **inferenční mechanismus**
- ▶ znalost **komunikační situace**

Roviny analýzy jazyka – příklad

rovina analýzy

příklad

pragmatická

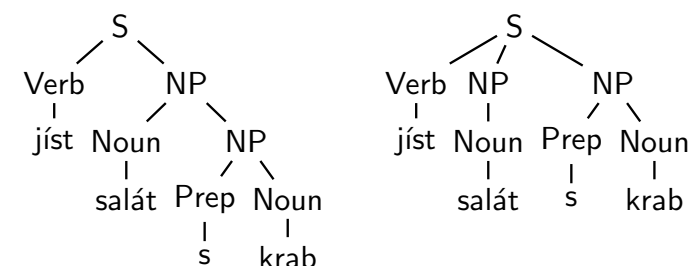
$Jíst(I_2, Salát, Tl_3) \wedge with(Salát, krab)$

sémantická

$Jíst(On, Salát, T_{min}) \wedge with(Salát, krab)$

syntaktická

$Jíst_{with}(On, Salát, krab, T_{min})$



morfologická

jíst–Verb3MSP, salát–Noun4IS, s–Prep7, krab–Noun7MS

fonetická

[j e d l s a l a : t s k r a b e m]

povrchová

“Jedl salát s krabem.”

Roviny analýzy jazyka – pokrač.

- ▶ **fonetická** – postihuje vztahy mezi zvuky používanými v (mluveném) jazyce, jejich skládání do slabik a slov
foném – nejmenší jednotka jazyka, která může **odlišit** význam nadřazených jednotek

kosit/nosit fonémy *k* a *n* odlišují dvě slova

často odpovídají *znakům* → vždy ale označují *zvuky*

- ▶ **morfologická** – interní struktura slov, skládání slov z menších jednotek
morfém – nejmenší jednotka, která může **nést** význam
pří-lež-it- pří – prefix (*blízko*)
-ost-n-ými: lež – lexikální kořen (*ležet*)
it – adjektivní derivační sufix (*ten, který*)
ost – substantivní derivační sufix (*ta skutečnost, že*)
n – adjektivní derivační sufix (*charakteristický pro*)
ými – gramatický afix (*instrumentál plurálu*)

Fonetika a fonologie

Fonologie:

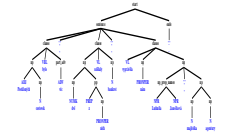
- ▶ **fonologický systém** jazykových zvuků v *určitém jazyce*
- ▶ pracuje s **gramatikou** řečových zvuků
- ▶ pomocí gramatických pravidel popisuje historické změny i současné alternace

Fonetika:

- ▶ studuje **produkcí, přenos a příjem** jazykových zvuků
- ▶ má klíčový význam např. pro oblast automatického **rozpoznávání** a **syntézy řeči**
- ▶ není tradičně chápána jako součást gramatiky jazyka

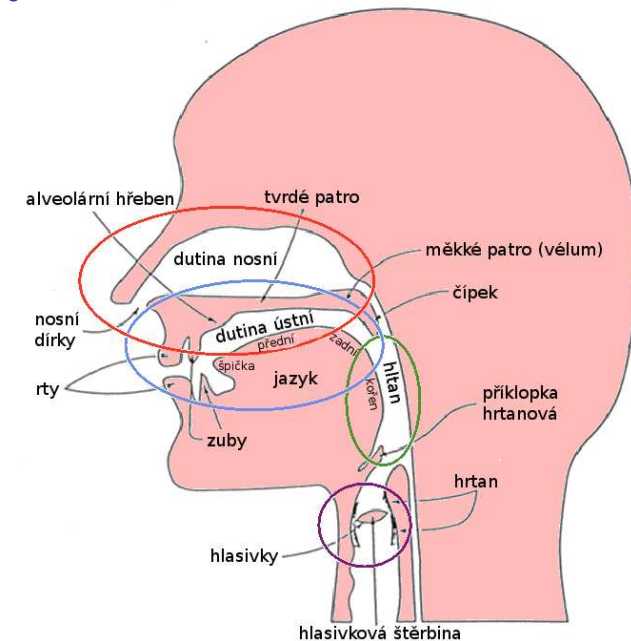
Roviny analýzy jazyka – pokrač.

- ▶ **syntaktická** – struktura větných frází popisuje, jak vypadá **gramaticky správná věta**, většinou pomocí **pravidel gramatiky**
syntaktický analyzátor – nástroj, který analyzuje vstup na základě gramatiky na výstup dává různé info, např. derivační stromy



- ▶ **sémantická** – význam výrazů přirozeného jazyka a jejich kombinací hodně závisí na zvolené **sémantické reprezentaci**
logická analýza věty – strukturní část sémantické analýzy
- ▶ **pragmatická** – zkoumá vztah mezi výrazy přirozeného jazyka a **kontextem** často se do ní řadí znalost **komunikační situace, základní ontologie a jazykových metaznalostí**

Kde vznikají jazykové zvuky?



Členění řečového proudu

řečový proud:

- ▶ nejsou mezery mezi slovy
- ▶ nejsou žádné izolované zvuky
- ▶ přesto všechny jazyky pracují s lingvistickými jednotkami jako separátními

oronym/orofón – fráze, které zní stejně/podobně, ale mají jiný obsah (≠ české *oronymum!*)

It's not easy to recognize speech.
It's not easy to wreck a nice beach.

Fonetická transkripce

- ▶ jeden z nejpoužívanějších **nástrojů fonetiky**
- ▶ **převod** řečového proudu do oddělených, lingvisticky významných **symbolických jednotek**
- ▶ používá se standardních **fonetických abeced** (viz dále)
- ▶ **široká** × **úzká** (broad/narrow) transkripce = převod *do fonémů/fónů*
- ▶ důvody pro tento převod: nedostatečnost písmenného zápisu, mezijazykové/krajové variace v písmenném zápisu
 - jedno písmeno → různý zvuk
 - vypít [v] / vpustit [f]
 - 'k' → 'c' v latinském *canis*, 'ch' v italském *Chianti*
 - 'č' → 'ch' v anglickém *cheat*, 'ci' v italském *ciao*
 - jeden zvuk → různá písmena
příp. jeden foném může být zaznamenán více písmeny
 - chovat [x] / shánět [x]
 - 'f': → 'f' v českém fyzika
→ 'gh' v anglickém *laugh*
→ 'ph' v řeckém *philosophia*

Fonetické jednotky

▶ foném (*phoneme*)

- ▶ základní jednotka **zvukového systému** jazyka
- ▶ foném je *abstraktní věc*, konkretizuje se pomocí *fónů* (viz dále)
- ▶ např. v **češtině** – 37 fonémů:

a, a:, b, ts, tS, d, d', dz, dZ, e, e:, f, g, h\, x, i, i:, j, k, l, m, n, n', o, o:, p, r, r', s, S, t, t', u, u:, v, z, Z

▶ fón (*phone*)

- ▶ **řečový zvuk** z hlediska jeho **fyzikálních charakteristik** (zvuková vlna určitého tvaru)
- ▶ bez zařazení k zvukovému systému jazyka
- ▶ jeden **foném** odpovídá **množině** fónů
- ▶ **alofón** určitého fonému = jeden z množiny fónů tohoto fonému
např. *nosit*, *banka*

Příklady dat pro českou transkripci pro MBROLA

- ▶ pravidla pro přepis do fonémů

```
CLASS SA [aáéěiioóuúýý] # samohlásky
CLASS ZPS [bdd'gvzžhCČ] # znělé párové souhlásky
CLASS NPS [ptt'kfsšHcč] # neznělé párové souhlásky
[[ dě ]] → d' e
[[ b ]] ( _|NPS|ZPS_ ) → p
[[ p ]] ZPS → b
```

- ▶ vstup pro MBROLu – text “shání tě těž muž”

```
_ 200 0 132          i: 93 0 114          S 81 0 114
z 57 0 115          t' 27 0 120          m 43 0 120
h 45                e 50 0 114          u 61
a: 137              t 31 0 120          S 110
n' 75 0 132         e: 102                #
```

- ▶ zvuková databáze *cz2* – 37 fonémů, 1442 difónů
nutné ručně “nařezat” všechny difóny

Fonetické abecedy IPA a SAMPA

IPA:

- ▶ *International Phonetic Alphabet*
- ▶ vznikla v roce 1886 v Paříži, od té doby mnoho revizí (poslední 1996)
- ▶ speciální znak pro vyjádření každého **fónu**
- ▶ mezinárodně **standardní zápis** – jsou k dispozici tabulky a fonty
- ▶ *Unicode* – speciální IPA znaky v rozsahu U+0250–02AD

SAMPA:

- ▶ *Speech Assessment Methods Phonetic Alphabet*
- ▶ vznikla v projektu SAM (Speech Assessment Methods) v letech 1987–89
- ▶ **strojově čitelná** fonetická abeceda
- ▶ <http://www.phon.ucl.ac.uk/home/sampa/>

IPA – souhlásky ve slovech

p	<u>plate</u> , <u>piece</u> , <u>spin</u> , <u>capital</u> , <u>stop</u> , <u>tramp</u>	s	<u>ceiling</u> , <u>slim</u> , <u>psychology</u> , <u>Pacific</u> , <u>nasty</u> , <u>pass</u>
t	<u>trip</u> , <u>time</u> , <u>winter</u> , <u>retire</u> , <u>wait</u> , <u>front</u>	z	<u>zoo</u> , <u>zipper</u> , <u>hazard</u> , <u>prison</u> , <u>cares</u> , <u>breeze</u>
k	<u>kite</u> , <u>climb</u> , <u>character</u> , <u>rocket</u> , <u>back</u> , <u>sink</u>	ʃ	<u>shore</u> , <u>sugar</u> , <u>nation</u> , <u>rash</u> , <u>Porche</u>
b	<u>bill</u> , <u>brush</u> , <u>sober</u> , <u>ramble</u> , <u>sob</u> , <u>bulb</u>	ʒ	(<u>genre</u>), <u>visual</u> , <u>measure</u> , <u>decision</u> , <u>massage</u>
d	<u>dark</u> , <u>drive</u> , <u>redden</u> , <u>ponder</u> , <u>head</u> , <u>hard</u>	h	<u>hat</u> , <u>who</u> , <u>ahead</u> , <u>perhaps</u>
g	<u>go</u> , <u>grease</u> , <u>rigor</u> , <u>anger</u> , <u>log</u> , <u>iceberg</u>	tʃ	<u>China</u> , <u>cheap</u> , <u>ritual</u> , <u>teaching</u> , <u>beach</u> , <u>punch</u>
m	<u>man</u> , <u>mile</u> , <u>remorse</u> , <u>ample</u> , <u>climb</u> , <u>harm</u>	dʒ	<u>jump</u> , <u>pidgeon</u> , <u>reject</u> , <u>individual</u> , <u>ridge</u> , <u>engine</u>
n	<u>nice</u> , <u>know</u> , <u>enough</u> , <u>cunning</u> , <u>sign</u> , <u>burn</u>	l	<u>light</u> , <u>look</u> , <u>pillow</u> , <u>applaud</u> , <u>salt</u> , <u>ball</u> , <u>girl</u>
ŋ	<u>finger</u> , <u>singer</u> , <u>drunk</u> , <u>rang</u> , <u>thing</u>	r	<u>real</u> , <u>row</u> , <u>around</u> , <u>part</u> , <u>care</u> , <u>hear</u>
θ	<u>thank</u> , <u>three</u> , <u>ether</u> , <u>panther</u> , <u>path</u> , <u>birth</u>	w	<u>wind</u> , <u>was</u> , <u>await</u> , <u>swim</u> , <u>queen</u>
ð	<u>then</u> , <u>these</u> , <u>feather</u> , <u>breathe</u>	j	<u>yes</u> , <u>use</u> , <u>beyond</u> , <u>beauty</u> , <u>punitive</u>
f	<u>fit</u> , <u>fly</u> , <u>effort</u> , <u>perform</u> , <u>enough</u> , <u>Ralph</u>		
v	<u>very</u> , <u>view</u> , <u>every</u> , <u>prevail</u> , <u>love</u> , <u>starye</u>		

IPA – souhlásky

v americké angličtině – *pulmonické* i *nepulmonické*

	labio-		alveolára				palatála		velára	glotála	
	labiála	dentála	dentála								
ploziva	p	b			t	d			k	g	
frikativa			f	v	θ	ð	s	z	ʃ	ʒ	h
afrikáta									tʃ	dʒ	
nazála		m				n				ŋ	
aproximanta						l					
laterální						r					
retroflexní											
koartikulovaná		w						j			

IPA – samohlásky

v americké angličtině

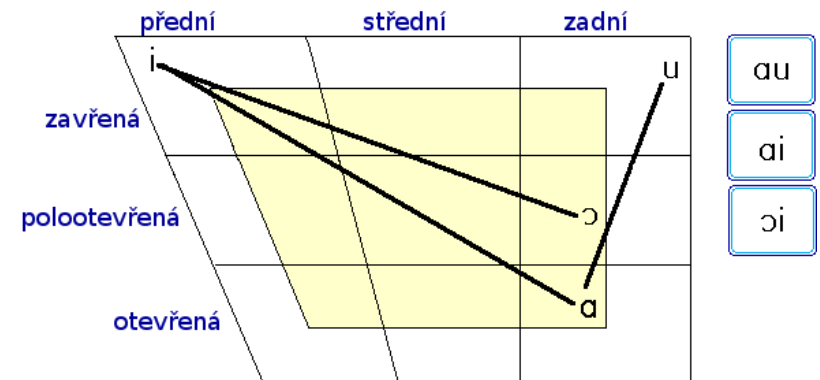
	přední	střední	zadní
zavřená	i		u
	ɪ		ʊ
polootevřená	e	ə	o
	ɛ	ʌ	ɔ
otevřená		æ	ɑ

IPA – samohlásky ve slovech

i	heed, beat, believe, people, scary	u	food, boot, pool, through, who, sewer
I	hid, bit, injure, resist, finish	U	hood, book, pull, put, would
e	hate, bait, great, they, say, neighbor	o	hole, boat, sew, know, so
ɛ	head, bet, friend, says, guest	ɔ	bought, law, wrong, stalk
æ	had, bat, laugh, calf, language	ɑ	pot, "la", stocking, father, rob
ə	above, around, sofa, police		
ʌ	bus, rush, under, other		

IPA – dvojhlásky

v americké angličtině



ai	find, high, aisle, quiet, ride
au	house, crown, around, flower, how
oi	boy, enjoy, Freud, avoid, join

Text-to-Speech systémy

- ▶ **syntéza řeči** – převod psaného textu na (digitální) zvuk
- ▶ TTS, *Text-to-Speech*
- ▶ dvě hlavní části
 1. **jazykový modul**, NLP modul
 - vstup = text
 - výstup = fonémy + prozodická informace
 - označována také jako TTP, *Text-to-Phoneme*
 2. **modul zpracování signálu**, DSP (Digital Signal Processing) modul
 - vstup = výstup z NLP modulu
 - výstup = zvukový soubor

Příklady TTS systémů

- ▶ české
 - **Epos** – z 90. let, Karlova univerzita a ČAV, nejlepší český open source
 - **Demosthenes** – FI MU Brno, laboratoř LSD
slabiková syntéza, základní prozodie
 - **ARTIC** (ARTificial Talker In Czech) – ZČU Plzeň, **DEMO**
obsahuje i "Talking head" vizuální část
 - **CS-Voice 97** – komerční, Frog Systems, pro Windows
- ▶ zahraniční
 - **Festival** – z Edinburghu, GPL, hodně jazyků, projekt Festival Czech
 - **MBROLA** – difónová syntéza MBR-PSOLA, řeší DSP část
Mikuláš Piňos, DP 2000 – česká DB pro MBROLu, text2phone
v Perlu
 - mnohé další – **HADIFIX**, **SVOX**, **Bell Labs**, **AT&T**, ...