Roviny analýzy jazyka. Fonetika

Aleš Horák

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

Obsah:

- Roviny analýzy jazyka
- Fonetika a fonologie

Struktura jazyka

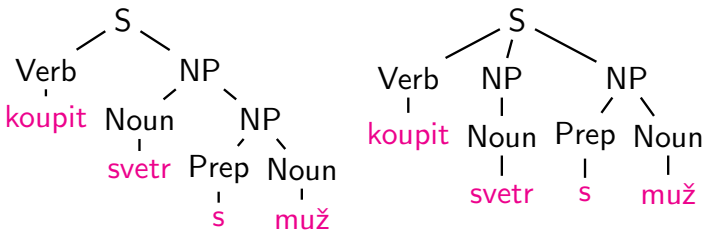
Struktura jazyka zahrnuje informace o:

- 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:

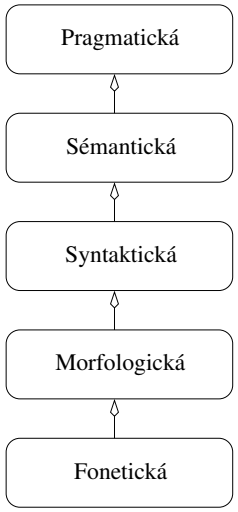
- obecnou (encyklopedickou) znalost světa (ontologie)
- inferenční mechanismus
- znalost komunikační situace

Roviny analýzy jazyka – příklad

rovina analýzy	**příklad**
pragmatická	$Koupit(I_2, Svetr, TI_3) \wedge accomp\_by(I_2, Muž)$
sémantická	$Koupit(Ona, Svetr, T_{min}) \wedge with(Svetr, muž)$ $Koupit(Ona, Svetr, T_{min}) \wedge accomp\_by(Ona, Muž)$
syntaktická	
morfologická	koupit–Verb3FSP, svetr–Noun4IS, s–Prep7, muž–Noun7MS
fonetická	[k o_u p i l a s v e t r s m u Z e m]
povrchová	"Koupila svetr s mužem."

Roviny analýzy jazyka

znalosti struktury jazyka jsou propojeny hierarchicky



jazykové roviny:

- fonetická
- morfologická
- syntaktická
- sémantická
- pragmatická

- kontextová
- znalost základní ontologie
- jazykové metaznalosti

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*

  NLP úkoly: např. syntéza řeči, rozpoznávání řeči, rozpoznávání emocí v hlase

Roviny analýzy jazyka – pokrač.

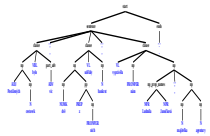
- morfologická – interní struktura slov, skládání slov z menších jednotek
  morfém – nejmenší jednotka, která může nést význam

  pří-lež-it-
  -ost-n-ými:
  - pří – prefix (*blízko*)
  - 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*)

  NLP úkoly: např. indexování textů, korektury pravopisu, analýza sentimentu, získávání informací, modelování tématu/stylu

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



  NLP úkoly: např. generování jazyka, extrakce informací, korektury gramatiky, extrakce termínů a klíčových frází

Roviny analýzy jazyka – pokrač.

- sémantická – význam výrazů přirozeného jazyka a jejich kombinací hodně závisí na zvolené sémantické reprezentaci
  logická analýza věty – strukturní část sémantické analýzy

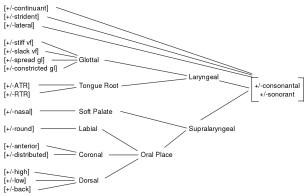
  NLP úkoly: např. strojový překlad, odpovídání na otázky, sumarizace textu

- pragmatická – zkoumá vztah mezi výrazy přirozeného jazyka a kontextem
  často se do ní řadí znalost komunikační situace, základní ontologie a jazykových metaznalostí

  NLP úkoly: např. porozumění textu, dialog člověk–stroj, zpřesněné verze úkolů z ostatních vrstev

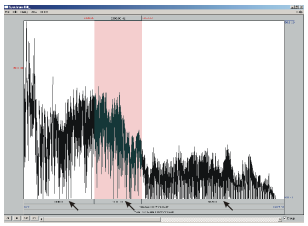
Fonetika a fonologie

Fonologie:



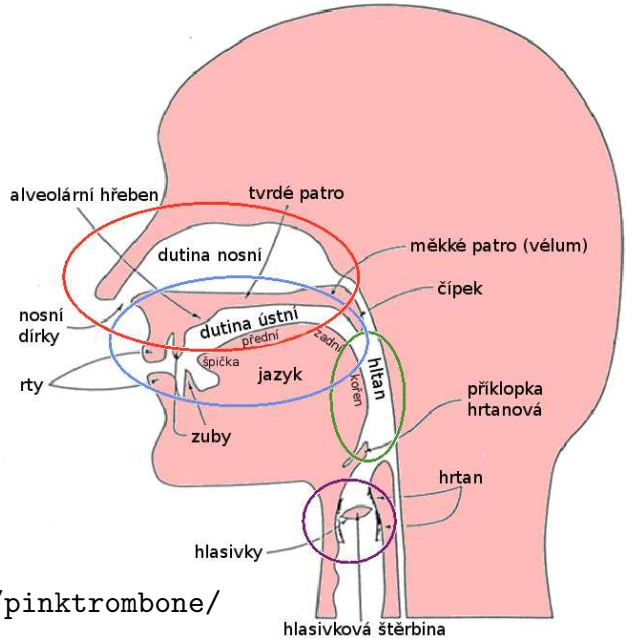
- fonologický systém jazykových zvuků v *určitém jazyce*
- pracuje s gramatikou řečových zvuků
- pomocí gramatických pravidel popisuje historické změny i současné alternace

Fonetika:



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

Kde vznikají jazykové zvuky?



`https://dood.al/pinktrombone/`

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 ($\neq$ české *oronymum*!)

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

v češtině např. *pohádce*/*po hádce*, *vesnu*/*ve snu*, *máti*/*má ti*
častější homofonní slova – *být*/*bít*, *výška*/*vížka*, *přeskáče*/*přezkáče*, *sběh*/*zběh*

Fonetické jednotky

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

Fonetická transkripce

- jeden z nejpoužívanějších nástrojů fonetiky
- převod řečového proudu do 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
    - spodoba znělosti v češtině: *vypít* [v] / *vpustit* [f]
    - krajové variace výslovnosti: *shánět* – moravská [z h], česká [s ch]
    - 'c' → [c] v latinském *Cicero*, [k] v *canis*, [č] v italském *ciao*
    - 'ch' → [ch] v čes. *chovat*, [č] v angl. *cheat*, [k] v it. *Chianti*
  - jeden zvuk → různá písmena (foném může být zaznamenán více písmeny)
    - [j] → 'j' v českém *jídlo*
      → 'y' v anglickém *yes*
      → 'ea' v anglickém *beautiful*
    - [f] → 'f' v českém *fyzika*
      → 'gh' v anglickém *laugh*
      → 'ph' v řeckém *philosophia*

Příklady dat pro českou transkripci pro MBROLA

- pravidla pro přepis do fonémů

```
CLASS SA    [aáeéěiíoóuúůyý]   # 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 MBROLA – text "shání tě též 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 několik revizí (poslední 1996, drobnosti pak 2005 a 2015)
- speciální znak pro vyjádření každého fónu
- mezinárodně standardní zápis – jsou k dispozici tabulky a fonty
- *Unicode* – speciální IPA znaky v rozsahu U+0250–02AD
  zápis např. `www.i2speak.com`

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

v americké angličtině – *pulmonické* i *nepulmonické*

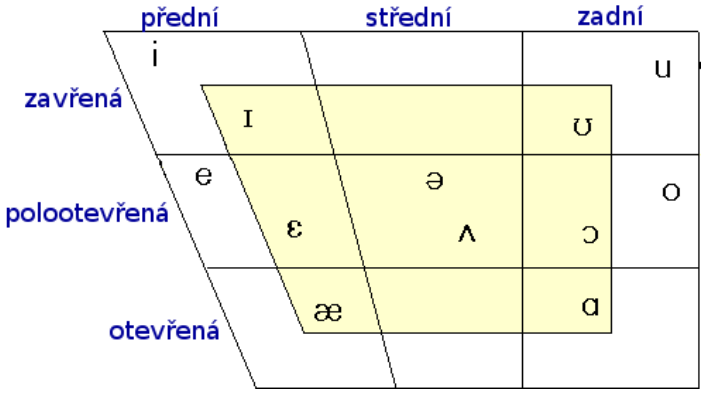
	labiála		labio-dentála		dentála		alveolára		palatála		velára		glotála	
ploziva	p	b					t	d			k	g		
frikativa			f	v	θ	ð	s	z	ʃ	ʒ			h	
afrikáta									tʃ	dʒ				
nazála		m						n				ŋ		
aproximanta								l						
laterální														
retroflexní								r						
koartikulovaná		w								j				

IPA – souhlásky ve slovech

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

IPA – samohlásky

v americké angličtině
vokalický čtyřúhelník

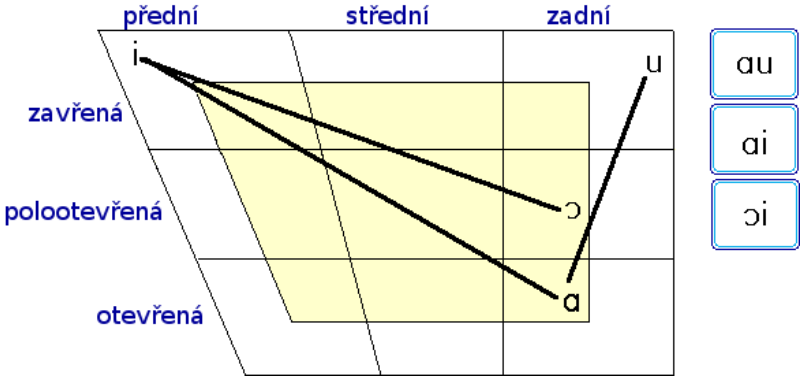


IPA – samohlásky ve slovech

i heed, beat, believe, people, scary
ɪ hid, bit, injure, resist, finish
e hate, bait, great, they, say, neighbor
ɛ head, bet, friend, says, guest
æ had, bat, laugh, calf, language
u food, boot, pool, through, who, sewer
ʊ hood, book, pull, put, would
o hole, boat, sew, know, so
ɔ bought, law, wrong, stalk
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
ɔi 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 se vztahem k češtině

- Epos – z 90. let, Karlova univerzita a ČAV, nejlepší český open source
- MBROLA – difónová syntéza MBR-PSOLA, řeší DSP část
  Mikuláš Piňos, DP 2000 – česká DB pro MBROLA, `text2phone` v Perlu
- 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
- Espeak – open source, formantová syntéza, včetně češtiny
- Festival – z Edinburghu, GPL, hodně jazyků, projekt Festival Czech