Morfologie, morfologická analýza

Aleš Horák

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

Obsah:

- Morfologie
- Morfologická analýza

Morfologie

- nauka o stavbě a tvorbě slov (v daném jazyce)
- morfém – nejmenší jednotka, která může nést význam

pří-lež-it-ost-n-ými

základní tvar = příležitostný
příd. jméno, rod muž. živ., neživ., žen. nebo stř., 7. pád, mn. č.

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*)

Základní lingvistické termíny v morfologii

- slovní druh – podstatné jméno (*substantivum*), přídavné jméno (*adjektivum*), sloveso (*verbum*), příslovce (*adverbium*), ...
- pád – *nominativ*, *genitiv*, *dativ*, *akuzativ*, *vokativ*, *lokál*, *instrumentál*
- číslo – *singulár*, *plurál*
- rod – 4 rody, mužský (*masculinum*) životný a neživotný (*animativní* a *inanimativní*), ženský (*femininum*) a střední (*neutrum*)
- slovotvorba – předpona (*prefix*), přípona (*sufix*), předpona nebo přípona (*afix*)
- základní tvar slova – *lemma* (mn. č. *lemmata*)
- ohýbání slov (*flexe*) – skloňování (*deklinace*) a časování (*konjugace*)
- odvozování – *derivování*

Dělení morfémů

dělení používané zejména v analytických jazycích (angličtina):

- morfémy obsahové (*content*) × funkční (*function*)
- morfémy volné (*free*) × vázané (*bound*)

dělení používané zejména ve flektivních jazycích (čeština):

- kořeny – nesamostatné morfémy nesoucí elementární lexikální významy
- afixy, které se dále dělí
  - podle funkce:
    - *gramatické*/*flektivní* – vyjadřují gramatické kategorie
    - *slovotvorné*/*derivační* – odvozování slov
  - podle postavení vzhledem ke kořeni:
    - *prefixy* – morfémy stojící před kořenovým morfémem (pod-, anti-, v-)
    - *sufixy* – morfémy připojované za kořenové morfémy (-ík, -izmus, ...)
    - *postfixy* – slovotvorné morfémy připojované až za gramatický sufix (kdosi, kohokoli, ...)
    - *circumfix* – morfémy připojované "kolem" základu, není v češtině
    - *infix*, *interfix* – morfémy vsazované dovnitř slova (mal-il-inký, velk-o-město, ...)

Procesy tvoření slov

dělení **morfologie** podle třech základních procesů tvoření slov:

- flektivní morfologie – popisuje strukturu slovních tvarů pomocí flexe (ohýbání – skloňování a časování)

1	pes	2	psa	3	psovi, psu	4	psa
5	pse	6	psovi, psu	7	psem		
1	*psové*, psi	2	psů	3	*psum*, psům	4	psy
5	*psové*, psi	6	*psách*, psech	7	psy, *psama*		

- derivativní (derivační) morfologie – zkoumá odvozování slov

  mýdlo: mydl-ář, mydl-ina, mýdel-ný, mydl-it, mýdél-ko

- kompozicionální (kompoziční) morfologie – zachycuje tvoření slov pomocí skládání

  ohni-vzdorný, pravdě-podobný, oka-mžik
  tlako-měr, vodo-pád, děje-pis
  samo-obsluha, malo-město, bylo-žravý

Derivační morfologie – vztah fundace

fundace – základní slovotvorný vztah

- slova neutvořená, prvotní, fundující – nemůžeme vysvětlit pomocí jiných slov jazyka

  voda, hlava, vejce

- slova utvořená, fundovaná – opírají se o slova základová

  trávník, růžový, učitel

- fundace – spojení slova základového se slovem utvořeným

  mladý $\rightarrow$ mladík

- slovotvorná řada – opakované odvození až k prvotnímu slovu

  rybníkářský $\leftarrow$ rybníkář $\leftarrow$ rybník $\leftarrow$ ryba

Derivační morfologie – vztah fundace

- slovotvorný svazek/hnízdo – souhrn slov fundovaných jedním slovem

  mýdlo → mydl-ář, mydl-ina, mýdel-ný, mydl-it, mýdél-ko

- slovotvorná čeleď – souhrn všech příbuzných slov (se stejným kořenem)

  les

  - pra-les → pra-les-ní
  - les-ní
    - lesn-ík → lesnic-ký → lesnic-tví
    - lesn-ice
    - nad-lesní
  - les-ík → lesíč-ek

Lexikální a gramatické kategorie

Morfologická analýza klasifikuje (značkuje, *tag*) slovní tvary jednotlivých kategorií (Part of Speech/PoS tags). Kategorie pro účely analýzy můžeme dělit na dvě skupiny:

- lexikální kategorie – pojmenovávají věci, akce, myšlenky
  podstatná jména, slovesa, přídavná jména, příslovce, ...
- gramatické kategorie – vyjadřují vztahy mezi ostatními větnými členy
  předložky, spojky, částice, anglické členy, ...

jazyky s
- jednoduchou morfologií (angličtina) – několik desítek kategorií (*POS – Part of Speech* – slovní druhy)
- bohatou morfologií – hierarchický systém, kde vedle základních slovních druhů určujeme nejrůznější subklasifikace (pád, číslo, rod, osoba, druhy příslovcí, ...) – celkově tisíce značek

Morfologická analýza

- rozpoznávání slovních tvarů
- nástroj se nazývá morfologický analyzátor (*Part-of-Speech/PoS tagger*)
- provádí lemmatizaci – přiřazuje k rozpoznaným slovním tvarům základní tvar (lemma)
- charakterizuje morfo-syntaktické vlastnosti nalezených slovních tvarů:

```
příležitostného
1. <s> příležitostn-ého (mladý GcAa)
   <l> příležitostný
   <c> adje      Man sg #4
   <c> adje      Man,Min,Neu sg #2
```

- kvalita morfologické analýzy ovlivňuje všechny následující analytické roviny

Morfologická analýza

Úkol morfologické analýzy zahrnuje 3 podúkoly:

- vypsat všechny možné analýzy – klasický morfologický analyzátor

```
<s> =svěž=í== (331-cizí)
    <l>svěží
    <c>k2eAgMnSc1d1 <c>k2eAgMnSc5d1 <c>k2eAgMnPc1d1 <c>k2eAgMnPc4d1
    <c>k2eAgInSc1d1 <c>k2eAgInSc4d1 <c>k2eAgInSc5d1 ...
```

- vybrat jednu nejpravděpodobnější analýzu – značkovač (tagger)

Svěží vánek zanesl do naší vesnice příchuť jara.

```
<s>
Svěží svěží k2eAgInSc1d1
vánek vánek k1gInSc1 ...
```

- analýzy pro neznámé slovo podle koncovky – "hádač" (guesser)

memorizovatelnými:

- ajka: –notfound
- guesser: `memorizovatelnými <l>memorizovatelný <c>k1gFnPc7`

Anglické gramatické morfémy

`-s`	3. osoba, jedn.č., přítomný čas
`-ed`	minulý čas
`-ing`	průběhový
`-en`	příčestí minulé trpné
`-s`	množné číslo
`-'s`	přivlastnění
`-er`	2. stupeň přídavného jména (komparativ)
`-est`	3. stupeň přídavného jména (superlativ)

Pro získání základního tvaru (pro indexování) často stačí *odsekávat koncovky* (*stemming*)

Automatické značkování

- Part-Of-Speech Tagging

  The/DT girls/NNs learned/VVD basic/AJ martial/AJ arts/NNs poses/NNs.

- učení z trénovacích dat
  - s dohledem (supervised) – vzorové texty i značky
  - bez dohledu (unsupervised) – pouze texty
  - s částečným dohledem (semi-supervised) – texty a výstup morf.analyzátoru (s pravděpodobnostmi)

Brillův značkovač

- učí se podle trénovacích dat:
  1. přiřaď nejčastější značku
  2. zkontroluj, kde jsou chyby (podle trénovacích dat)
  3. ohodnoť pravidla pro opravu chyb → vyber nelepší → oprav zpětně chybné značky
  4. opakuj, dokud se daří odvozovat dobrá pravidla
- používá učení založené na transformacích (*transformation-based learning*)
- analogie – malování obrazu: nejprve pozadí a pak přes něj stále drobnější detaily
- značkuje 36 různých POS značek
- úspěšnost – přes 90 %

Brillův značkovač – příklad

věta:	podle frekvence:	P1:	P2:	správně (zlatý standard):
The	at			at
President	nn-tl			nn-tl
said	vbd			vbd
he	pps			pps
will	md			md
ask	vb			vb
Congress	np			np
to	to			to
increase	nn	vb		vb
grants	nns			nns
to	to	to	in	in
states	nns			nns
for	in			in
vocational	jj			jj
rehabilitation	nn			nn
.	.			.

P1: Replace `nn` with `vb` when the previous word is `to`
P2: Replace `to` with `in` when the next tag is `nns`

Brillův značkovač – příklad

Loading tagged data...
Training unigram tagger: [accuracy: 0.820940]
Training Brill tagger on 37168 tokens...

Iteration 1: 1482 errors; ranking 23989 rules;
Found: "Replace POS with VBZ if the preceding word is tagged PRP"
Apply: [changed 39 tags: 39 correct; 0 incorrect]

Iteration 2: 1443 errors; ranking 23662 rules;
Found: "Replace VBP with VB if one of the 3 preceding words is tagged MD"
Apply: [changed 36 tags: 36 correct; 0 incorrect]

Iteration 3: 1407 errors; ranking 23308 rules;
Found: "Replace VBP with VB if the preceding word is tagged TO"
Apply: [changed 24 tags: 23 correct; 1 incorrect]
...
Iteration 21: 1128 errors; ranking 20569 rules;
Found: "Replace VBD with VBN if the preceding word is tagged VBD"
[insufficient improvement; stopping]

Brill accuracy: 0.835145

Algoritmický popis české formální morfologie

v češtině nestačí pravidla podle obecných morfémů – je potřebné mít lexikon, který ke každému *kmenu* obsahuje jeho přiřazení ke *vzoru*

morfologické (tvaroslovné) paradigma – soubor tvarů ohebného slova vyjadřující systém jeho mluvnických kategorií
vzor – reprezentace tvaroslovného paradigmatu paradigmatem určitého konkrétního slova

Algoritmický popis:

1. definice koncovkových množin
2. definice vzorů prostřednictvím vzorových slov rozdělených na:
   - neměnná část vzorového slova – kmenový základ
   - proměnlivé části vzorového slova – intersegmenty
   - koncovkové množiny obsahující utříděné seznamy všech přípustných koncovek vzorového slova spolu s jejich gramatickými významy

   popis vzoru = formální pravidlo, které specifikuje přípustné kombinace těchto komponent (segmentů) ohebného slova

Formát české morfologické databáze

slovník = `lemma:vzor|poznámka`

```
Luděk:Luděk|180.1
Vladěk:Luděk|180.1
hlemýžděk:Luděk|180.1
```

koncovkové množiny

```
=rs-mluv-S204          =rs-mluv-S386
  {_, k1gMnSc1}          {ů, k1gMnPc2}
=rs-mluv-S99           =rs-mluv-S499
  {i, k1gMnPc1}          {ovi, k1gMnSc3}
=rs-mluv-S102          {ovi, k1gMnSc6}
  {i, k1gMnPc5}        ...
```

vzor

```
+Luděk
    <děk> rs-mluv-S204
    <d'c> rs-mluv-S99, rs-mluv-S102, rs-mluv-S385
    <d'k> rs-mluv-S386, rs-mluv-S499, rs-mluv-S460,
          rs-mluv-konc12, rs-mluv-S510, rs-mluv-S74, rs-mluv-S71,
          rs-mluv-S294, rs-mluv-S521, rs-mluv-S522, rs-mluv-S163,
          rs-mluv-S171, rs-mluv-S299, rs-mluv-konc08
```

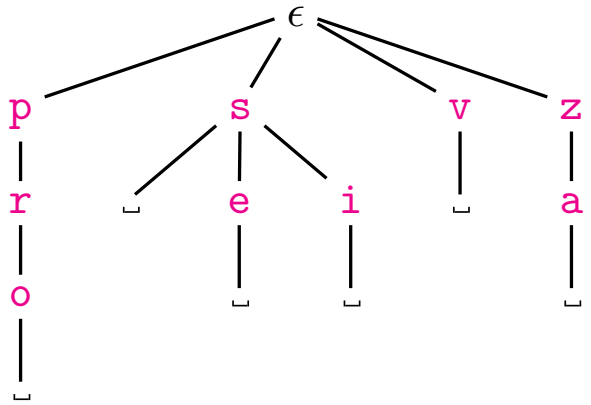

Segmentace slova pro potřeby algoritmického popisu

- segmentace od začátku slova
  a) segmenty se snadno formalizovatelným výskytem vázaným gramaticky:
     - negativní prefix `ne-`
     - superlativní prefix `nej-`
     - futurální slovesný prefix `po-`
  b) segmenty s nesnadno formalizovatelným výskytem vázáným sémanticky:
     - prefixy
     - první členy kompozit
     - prefixy `ni-`, `ně-` zájmen neurčitých a záporných
- segmentace od konce slova
  a) rozdělení slovního tvaru na kmen a koncovku
  b) další segmentace kmene na kmenový základ a intersegment

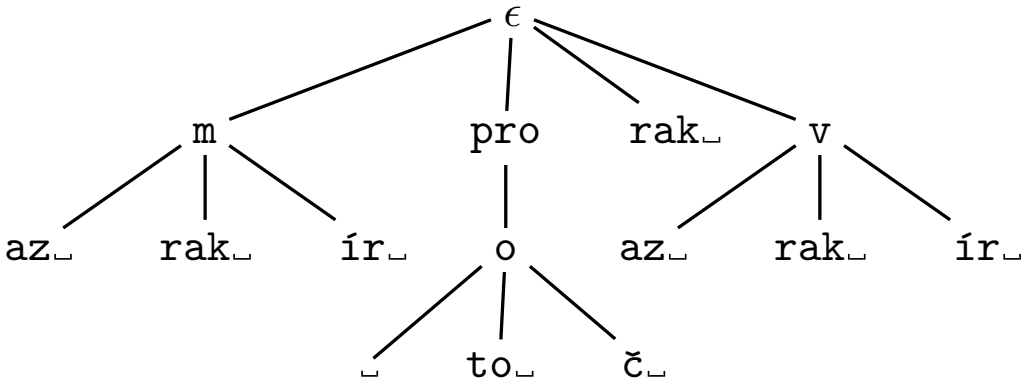
Efektivní implementace morfologického lexikonu – trie

struktura trie:

- uspořádaný strom nad danou abecedou A
- v každém uzlu je různé písmeno z abecedy A
- klíč je v trie uložen jako cesta od kořene
- výhody:
  - sdílení společných prefixů
  - v každém případě nalezení nejdelšího shodného prefixu



Eliminace cest v trie



Jiná efektivní implementace ML – konečný automat

- původně BP, Radovan Štancel, 2005 – doplňování diakritiky
- použití mírně pozměněných volně dostupných knihoven pro práci s KA od Jana Daciuka – FSA library
- vstupní data se generují ze slovníku ajky převedeného do tvaru "`slovo<TAB>lemma<TAB>značka`" (cca 33 mil. řádků)

```
Abcházce    Abcházec   k1gMnPc4
Abcházce    Abcházec   k1gMnSc2
Abcházce    Abcházec   k1gMnSc4
Abcházcem   Abcházec   k1gMnSc7
Abcházci    Abcházec   k1gMnPc1
Abcházci    Abcházec   k1gMnPc5
Abcházci    Abcházec   k1gMnPc7
Abcházci    Abcházec   k1gMnSc3
Abcházci    Abcházec   k1gMnSc6
...
```


Jiná efektivní implementace ML – konečný automat

- data se dále upravují pro KA – `slovo+zkr.lemma+značky`:

```
Abcházce+ACec+k1gMnPc4, k1gMnSc2, k1gMnSc4
Abcházcem+ADec+k1gMnSc7
Abcházci+ACec+k1gMnPc1, k1gMnPc5, k1gMnPc7, k1gMnSc3, ...
...
```

- v lemmatu – 1. písmeno je počet znaků, které se odtrhnou jako předpona, 2. písmeno je počet znaků, které se trhají od konce, a ostatní znaky se přidají
- tím se sníží počet řádků na 6.7 mil. řádků, ze kterých se přímo generuje (a minimalizuje) konečný automat
- výsledný slovník má 4.3 MB
- rychlost je cca o 1/4 lepší než u trie, velikost řádově srovnatelná

České morfologické analyzátory

- ajka
  - Radek Sedláček, FI MU Brno
  - `http://nlp.fi.muni.cz/projekty/ajka/`
  - značky jsou řetězce dvojic atribut–hodnota
  - napsaný v C
  - využívá struktury trie
  - 390 000 základních tvarů, 6 300 000 různých slovních tvarů, 15 000 různých značek, slovník 3.13 MB
  - rychlost analýzy – cca 18 000 slov/s
  - v současnosti nový nástroj majka od Pavla Šmerka, na principu konečných automatů, s novým mechanizmem vzorů
- pražský morfologický analyzátor
  - Barbora Hladká, Jan Hajič a jeho tým, ÚFAL MFF UK Praha
  - `http://ufal.mff.cuni.cz/czech-tagging/`
  - používá poziční značky
  - "free" část napsaná v Perlu, menší slovník (cca 76 000 základních tvarů, 6 000 koncovek)

Pražský morfologický analyzátor – poziční značky

pozice	kategorie	anglicky	česky
1	POS	Part of Speech	Slovní druh
2	SUBPOS	Detailed Part of Speech	Slovní poddruh
3	GENDER	Agreement Gender	Rod
4	NUMBER	Agreement Number	Číslo
5	CASE	Case	Pád
6	POSSGENDER	Possessor's Gender	Rod vlastníka
7	POSSNUMBER	Possessor's Number	Číslo vlastníka
8	PERSON	Person	Osoba
9	TENSE	Tense	Čas
10	GRADE	Degree of Comparison	Stupeň
11	NEGATION	Negation (by prefix)	Negace
12	VOICE	Voice	Slovesný rod
13	RESERVE1	Reserved for future use	Rezerva
14	RESERVE2	Reserved for future use	Rezerva
15	VAR	Variant, Style, Register	Varianta, styl

Pražský morfologický analyzátor – příklad

- vstup:

  Prezident rezignoval na svou funkci.

- výstup:

```
<csts>
<f cap>Prezident<MMl>prezident<MMt>NNMS1-----A----
<f>rezignoval<MMl>rezignovat_:T<MMt>VpYS---XR-AA---
<f>na<MMl>na<MMt>RR--4----------<MMt>RR--6----------
<f>svou<MMl>svůj-1_^(přivlast.)<MMt>P8FS4---------1
   <MMt>P8FS7---------1
<f>funkci<MMl>funkce<MMt>NNFS3-----A----
   <MMt>NNFS4-----A----<MMt>NNFS6-----A----
<D>
<d>.<MMl>.<MMt>Z:-------------
</csts>
```


Značky morfologického analyzátoru ajka

značka = řetězec dvojic *atributHodnota*: `k1gNnSc3`

`k`	slovní druh	`1` – podst. jméno, `2` – př. jméno, ...
`g`	rod	`M` – muž. životný, `I` – muž. neživotný, ...
`n`	číslo	`S` – jednotné, `P` – množné, `D` – duál
`c`	pád	1, 2, ..., 7
`p`	osoba	1, 2, 3
`m`	slovesný způsob	`F` – infinitiv, `R` – imperativ, ...
`a`	slovesný vid	`P` – dokonavý, `I` – nedokonavý
`t`	typ příslovcí	`T` – času, `L` – místa, `M` – způsobu, ...
`x`	typ spojky	`C` – souřadící, `S` – podřadící

Morfologický analyzátor ajka – příklad

- dávkově

```
Prezident <l>prezident <c>k1gMnSc1
rezignoval <l>rezignovat <c>k5eApMnStMmPaI <c>k5eApInStMmPaI
na <l>na <c>k7c4 <c>k7c6
svou <l>svůj <c>k3xOgFnSc4p3 <c>k3xOgFnSc7p3
funkci <l>funkce <c>k1gFnSc3 <c>k1gFnSc6 <c>k1gFnSc4
```

- interaktivně

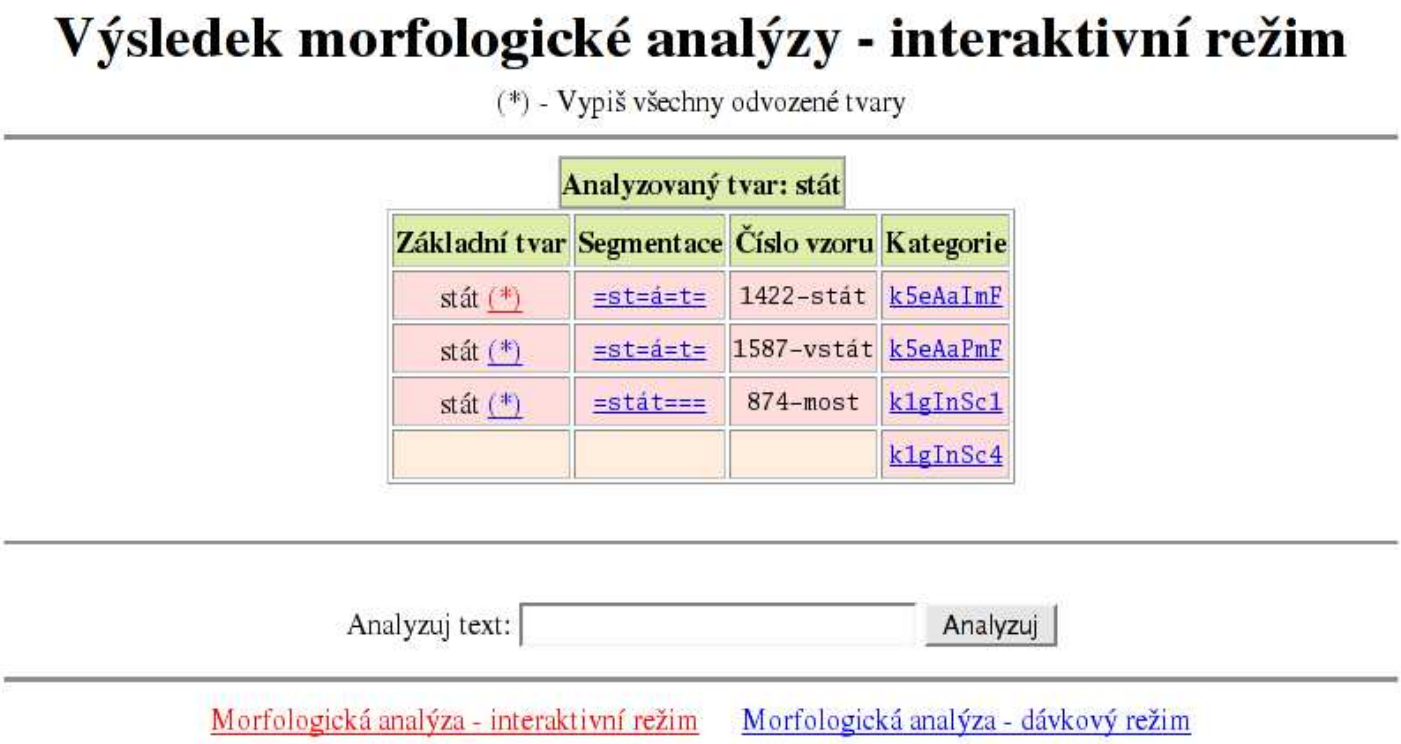
```
<s> ne=snesiteln=ého== (1023)
<l>snesitelný
<c>k2eNgMnSc2d1
<c>k2eNgMnSc4d1 ...
```

- všechny tvary (`ajka -a`)

```
<s> =p=es== (1148)
<l>pes
<c>k1gMnSc1
pes psům psů psovi psem psa psu psy psech pse psi psové
```


Morfologický analyzátor ajka – webové rozhraní

http://nlp.fi.muni.cz/projekty/wwwajka/



Výsledek morfologické analýzy - interaktivní režim

(*) - Vypiš všechny odvozené tvary

Analyzovaný tvar: stát			
Základní tvar	Segmentace	Číslo vzoru	Kategorie
stát (*)	=st=á=t=	1422-stát	k5eAaImF
stát (*)	=st=á=t=	1587-vstát	k5eAaPmF
stát (*)	=stát===	874-most	k1gInSc1
			k1gInSc4

Analyzuj text: Analyzuj

Morfologická analýza - interaktivní režim Morfologická analýza - dávkový režim