PLIN063_1
Algoritmický popis morfologie

osolsobe@phil.muni.cz

Osnova

- Přístup jazykovědce a informatika
- Fonetika a grafika
- Stabilní a proměnlivé
- Exaktní definice

Lingvistika a NLP

- Na střední škole se učí, že úkolem morfologické analýzy slova je určit morfologické kategorie danému slovu v textu příslušné.
- Pro člověka je tato definice přijatelná, a koneckonců každý z nás na oné střední škole nakonec nějak uspěl.
- Při počítačovém zpracování je však situaci třeba definovat a popsat mnohem přesněji.
- Především je třeba jasně rozlišovat mezi morfologickou kategorií a její hodnotou.

Přístup jazykovědce a informatika

- Segmentace slovního tvaru (lexikální kořeny, afixy, typy afixů)
- Terminologie
- Definice termínů
- Příklady (rozdíl mezi koncovkou a koncovým řetězcem, prefixem a iniciálním řetězcem)

Slovo a jeho tvary

- Bohatství tvaroslovného systému (systémové slovo/textové slovo)
- Vymezení tvaroslovného systému (Lemma/Word)
- Lemmatizace je závislá na tradici konkrétního jazyka i na rovině jazyka, kterou právě analyzujeme
- Technická řešení (nemusí být u všech nástrojů NLP identická)

Jak poznáme, co k sobě patří a co ne?

- *přípravku, přípravky, přípravek, přípravkem, přípravka*
- *správec, správce, správci, správcem*
- *buřt, buřtu, buřta, buřtem, buřtovi*
- *koblih, koblihu, kobliha, koblihem, koblihou*
- *myslím, myslíš, mysli, myslel, myslit, myslet*
- *citron, citrónu, citronem, citróny*
- *filozof, filosofa, filozofem, filosofovi*

Mezi jazyky existují značné rozdíly v bohatství a v pojetí tvaroslovného systému

	word (lowercase)	lempos (lowercase)	tag	Frekvence	Items: 9 || Total frequency: 27,221
P | N	love	love-n	NN	13,113	
P | N	love	love-v	VVP	4,346	
P | N	love	love-v	VV	3,518	
P | N	loved	love-v	VVD	3,192	
P | N	loved	love-v	VVN	1,376	
P | N	loves	love-v	VVZ	1,226	
P | N	loving	love-v	VVG	327	
P | N	loves	love-n	NNS	89	
P | N	love's	love-n	NNZ	34	

Značky pro slovesa

VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/present participle	taking
VVN	verb, past participle	taken
VVP	verb, present, not 3rd person	take
VVZ	verb, 3rd person sing. present	takes

Značky pro substantiva

NN	noun, singular or mass	table
NNS	noun plural	tables
NNSZ	possessive noun plural	people's, women's
NNZ	possessive noun, singular or mass	year's, world's
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
NPSZ	possesive proper noun, plural	Boys', Workers'
NPZ	possesive noun, singular	Britain's, God's

láska - substantivum

Celkem: 15 položek (1 stránka)

	Filter	lc [lowercase word]	pos [part of speech]	tag	Freq	
1	p / n	láska	N	NNFS1-----A-----	5860	
2	p / n	lásky	N	NNFS2-----A-----	5089	
3	p / n	lásku	N	NNFS4-----A-----	4057	
4	p / n	láskou	N	NNFS7-----A-----	2302	
5	p / n	lásce	N	NNFS6-----A-----	1801	
6	p / n	lásce	N	NNFS3-----A-----	705	
7	p / n	lásko	N	NNFS5-----A-----	621	
8	p / n	lásky	N	NNFP4-----A-----	186	
9	p / n	lásky	N	NNFP1-----A-----	162	
10	p / n	lásek	N	NNFP2-----A-----	143	
11	p / n	láskách	N	NNFP6-----A-----	54	
12	p / n	láskami	N	NNFP7-----A-----	31	
13	p / n	láskám	N	NNFP3-----A-----	17	
14	p / n	motivy–láska	N	NNFS1-----A-----	1	
15	p / n	láskama	N	NNFP7-----A---6-	1	

milovat - sloveso

Celkem: 53 položek (2 stránky)

	Filter	lc [lowercase word]	pos [part of speech]	tag	Freq	
1	p / n	miluje	V	VB-S---3P-AA---I	2811	
2	p / n	miluju	V	VB-S---1P-AA---I	2134	
3	p / n	milovat	V	Vf--------A----I	2104	
4	p / n	milovala	V	VpFS----R-AA---I	1666	
5	p / n	miloval	V	VpIS----R-AA---I	1251	
6	p / n	milují	V	VB-P---3P-AA---I	1174	
7	p / n	miluji	V	VB-S---1P-AA--1I	1120	
8	p / n	miloval	V	VpMS----R-AA---I	945	
9	p / n	milovali	V	VpMP----R-AA---I	818	
10	p / n	miluješ	V	VB-S---2P-AA---I	449	
11	p / n	milujeme	V	VB-P---1P-AA---I	372	
12	p / n	milujete	V	VB-P---2P-AA---I	299	
13	p / n	nemiluje	V	VB-S---3P-NA---I	219	
14	p / n	nemilovala	V	VpFS----R-NA---I	127	
15	p / n	milován	V	VsIS------AP---I	118	
16	p / n	milovaly	V	VpIP----R-AA---I	114	
17	p / n	nemiloval	V	VpMS----R-NA---I	97	
18	p / n	miluj	V	Vi-S---2--A----I	97	
19	p / n	nemiluju	V	VB-S---1P-NA---I	90	
20	p / n	nemiloval	V	VpIS----R-NA---I	90	

milulující - adjektivum

Celkem: 49 položek (1 stránka)

	Filter	lc [lowercase word]	pos [part of speech]	tag	Freq
1	p / n	milující	A	AGFS1-----A-----	114
2	p / n	milující	A	AGMS1-----A-----	107
3	p / n	milující	A	AGFS4-----A-----	78
4	p / n	milující	A	AGFS2-----A-----	50
5	p / n	milujícího	A	AGMS4-----A-----	47
6	p / n	milující	A	AGMP1-----A-----	39
7	p / n	milujícího	A	AGMS2-----A-----	34
8	p / n	milující	A	AGNS1-----A-----	21
9	p / n	milující	A	AGFP1-----A-----	20
10	p / n	milující	A	AGIS1-----A-----	19
11	p / n	milujícím	A	AGMS7-----A-----	19
12	p / n	milující	A	AGFP4-----A-----	17
13	p / n	milující	A	AGIS4-----A-----	16
14	p / n	milujícího	A	AGIS2-----A-----	14
15	p / n	milující	A	AGFS7-----A-----	13
16	p / n	milujících	A	AGMP2-----A-----	13
17	p / n	milujícím	A	AGIS7-----A-----	12
18	p / n	milujících	A	AGFP2-----A-----	12
19	p / n	milující	A	AGFS3-----A-----	10
20	p / n	milujícího	A	AGNS2-----A-----	10

Společné a odlišné

- Forma
- Funkce
- Význam

Formy + významy + funkce

rod: ž.

	jednotné číslo	množné číslo
1. pád	škola	školy
2. pád	školy	škol
3. pád	škole	školám
4. pád	školu	školy
5. pád	školo	školy
6. pád	škole	školách
7. pád	školou	školami

rod: s.

	jednotné číslo	množné číslo
1. pád	školství	školství[2]
2. pád	školství	školství
3. pád	školství	školstvím
4. pád	školství	školství
5. pád	školství	školství
6. pád	školství	školstvích
7. pád	školstvím	školstvími

rod: s.

	jednotné číslo	množné číslo
1. pád	školení	školení
2. pád	školení	školení
3. pád	školení	školením
4. pád	školení	školení
5. pád	školení	školení
6. pád	školení	školeních
7. pád	školením	školeními

	jednotné číslo	množné číslo
1. osoba	školím	školíme
2. osoba	školíš	školíte
3. osoba	školí	školí
rozkazovací způsob	škol[2]	školte
příčestí činné	školil	
příčestí trpné	školen	
přechodník přítomný, m.	škole[3]	školíce
přechodník přítomný, ž. + s.	školíc	
verbální substantivum	školení	

Tvary substantiv *slon, vlk, pes, papoušek*

rod: m. živ.

	jednotné číslo	množné číslo
1. pád	slon	sloni[4]
2. pád	slona	slonů
3. pád	slonovi, slonu[2]	slonům
4. pád	slona	slony
5. pád	slone[3]	sloni
6. pád	slonovi, slonu[2]	slonech
7. pád	slonem	slony

rod: m. živ.

	jednotné číslo	množné číslo
1. pád	vlk	vlci[4]
2. pád	vlka	vlků
3. pád	vlkovi, vlku[2]	vlkům
4. pád	vlka	vlky
5. pád	vlku[3]	vlci
6. pád	vlkovi, vlku[2]	vlcích[5]
7. pád	vlkem	vlky

rod: m. živ.

	jednotné číslo	množné číslo
1. pád	pes	psi, psové[4]
2. pád	psa	psů
3. pád	psovi, psu[2]	psům
4. pád	psa	psy
5. pád	pse[3]	psi, psové
6. pád	psovi, psu[2]	psech[5]
7. pád	psem	psy

rod: m. živ.

	jednotné číslo	množné číslo
1. pád	papoušek	papoušci, papouškové[4]
2. pád	papouška	papoušků
3. pád	papouškovi, papoušku[2]	papouškům
4. pád	papouška	papoušky
5. pád	papoušku[3]	papoušci, papouškové
6. pád	papouškovi, papoušku[2]	papoušcích[5]
7. pád	papouškem	papoušky

Porovnejte zápis dotazu v jazyce cql na slovní tvary slov tradičně řazených k jedinému vzoru

- Tvary substantiv *slon, vlk, pes, papoušek*
- **Dotaz**: slon([aueiůy]?|em|ům|ech|ov[ié])
- **Dotaz**: vlk([auůy]?|em|ům|ov[ié])|vlc(i|ích)
- **Dotaz**: pes|ps([aueiůy]|em|ům|ech|ov[ié])
- **Dotaz**: papoušek|papoušk([auůy]|em|ům|ov[ié])|papoušc(i|ích)

slon/vlk

Celkem: 12 položek (1 stránka)

	Filter	word	Freq
1	p / n	slon	421
2	p / n	sloni	336
3	p / n	slonů	280
4	p / n	slona	274
5	p / n	slony	226
6	p / n	slonové	52
7	p / n	slonem	44
8	p / n	slonovi	35
9	p / n	slonům	32
10	p / n	slonech	29
11	p / n	slonu	17
12	p / n	slone	5

Celkem: 11 položek (1 stránka)

	Filter	word	Freq
1	p / n	vlk	1154
2	p / n	vlka	614
3	p / n	vlci	590
4	p / n	vlků	443
5	p / n	vlky	253
6	p / n	vlkem	159
7	p / n	vlkovi	70
8	p / n	vlkům	66
9	p / n	vlku	53
10	p / n	vlcích	21
11	p / n	vlkové	13

pes/papoušek

Celkem: 12 položek (1 stránka)

	Filter	word	Freq
1	p / n	pes	5669
2	p / n	psa	4742
3	p / n	psi	2194
4	p / n	psy	2070
5	p / n	psů	1774
6	p / n	psem	993
7	p / n	psovi	577
8	p / n	psům	321
9	p / n	psu	215
10	p / n	psech	103
11	p / n	pse	50
12	p / n	psové	6

Celkem: 10 položek (1 stránka)

	Filter	word	Freq
1	p / n	papoušek	259
2	p / n	papouška	171
3	p / n	papoušků	124
4	p / n	papoušci	124
5	p / n	papoušky	78
6	p / n	papouškem	36
7	p / n	papouškovi	23
8	p / n	papouškům	10
9	p / n	papoušcích	8
10	p / n	papouškové	7

Popište zjištěné nesrovnalosti aparátem lingvistických termínů

- kmen
- Koncovka
- morfémový šev
- morfologická nula
- alomorfie
- hláskové alternace
- variantní koncovka

Strojový popis vzorů

- +slon {pán Ea}
- <> V1,V13X,VOVE,VVE,VZ1,VI
- <ův> PRIVL1X
- <ov> PRIVL1
- +pes {pán S}
- <> V13X
- +ps {pán S}
- <> V1,VOVE,VVE,VZ1,VI
- <ův> PRIVL1X
- <ov> PRIVL1

Strojový popis vzorů

- +vl {pán Fa}
- <k> V1,V13X,VOVE,VVU
- <kův> PRIVL1X
- <kov> PRIVL1
- <c> VI,VQ1
- +medvíd {pán Fb}
- <ek> V13X
- <k> V1,VOVE,VVU
- <kův> PRIVL1X
- <kov> PRIVL1
- <c> VI,VQ1

Segmentace slovního tvaru

- kmenový základ
- Intersegment
- Koncovkové množiny

Mnemotechnika

- Vztah ke klasickým vzorům (1=pán, 3=muž, …)
- Rozčlenění na koncovkové množiny podle n-tic koncovek se společnými vlastnostmi

https://nlp.fi.muni.cz/projekty/wwwajka/WwwAjkaSkripty/morph.cgi?jazyk=0

Výsledek morfologické analýzy – interaktivní režim

(*) - Vypiš všechny odvozené tvary

Analyzovaný tvar: papoušek			
Základní tvar	Segmentace	Číslo vzoru	Kategorie
papoušek (*)	=papouš=ek==	847-medvídek	k1gMnSc1

Odvozené tvary ke slovu "papoušek"

Rod: Mužský životný			
Pád	Singulár	Pád	Plurál
1	papoušek	1	papouškové, papoušci
2	papouška	2	papoušků
3	papoušku, papouškovi	3	papouškum, papouškům
4	papouška	4	papoušky
5	papoušku	5	papoušci, papouškové
6	papoušku, papouškovi	6	papouščích, papouškách
7	papouškem	7	papoušky, papouškama

Fonetika a grafika

- Porovnejte zápis dotazu v jazyce cql na slovní tvary slov tradičně řazených k jedinému vzoru
- Tvary substantiv *muž* a *choť*
- Tvary substantiv *růže* a *vůně*
- Tvary substantiv *kníže* a *hrabě*

Analyzovaný tvar: muž

Základní tvar	Segmentace	Číslo vzoru	Kategorie
muž (*)	=muž===	889-muž	k1gMnSc1

Analyzovaný tvar: choť

Základní tvar	Segmentace	Číslo vzoru	Kategorie
choť (*)	=cho=ť==	319-choť	k1gFnSc1
			k1gFnSc4
choť (*)	=cho=ť==	1704-zeť	k1gMnSc1

NLP jako aplikační oblast exaktního popisu jazyka

- Co víme o hláskových alternacích ve flexi
- Umíme formulovat pravidla?
- Víme, jak se vypořádat s výjimkami?

Pravidla distribuce variantních koncovek

- Víme, jaká pravidla platí?
- Umíme je formálně vyjádřit?
- Umíme ověřit jejich platnost/ nalézt výjimky?

Pravidla výskytu hláskových alternací

- Víme na čem závisí?
- Umíme to zjišťovat?

Výskyt hláskových alternací

- Je nějaké omezení místa alternace?
- Lze místo alternace formálně popsat?
- Potřebujeme morfémovou segmentaci, nebo lépe – v čem nám může pomoci?

Příklady - cvičení

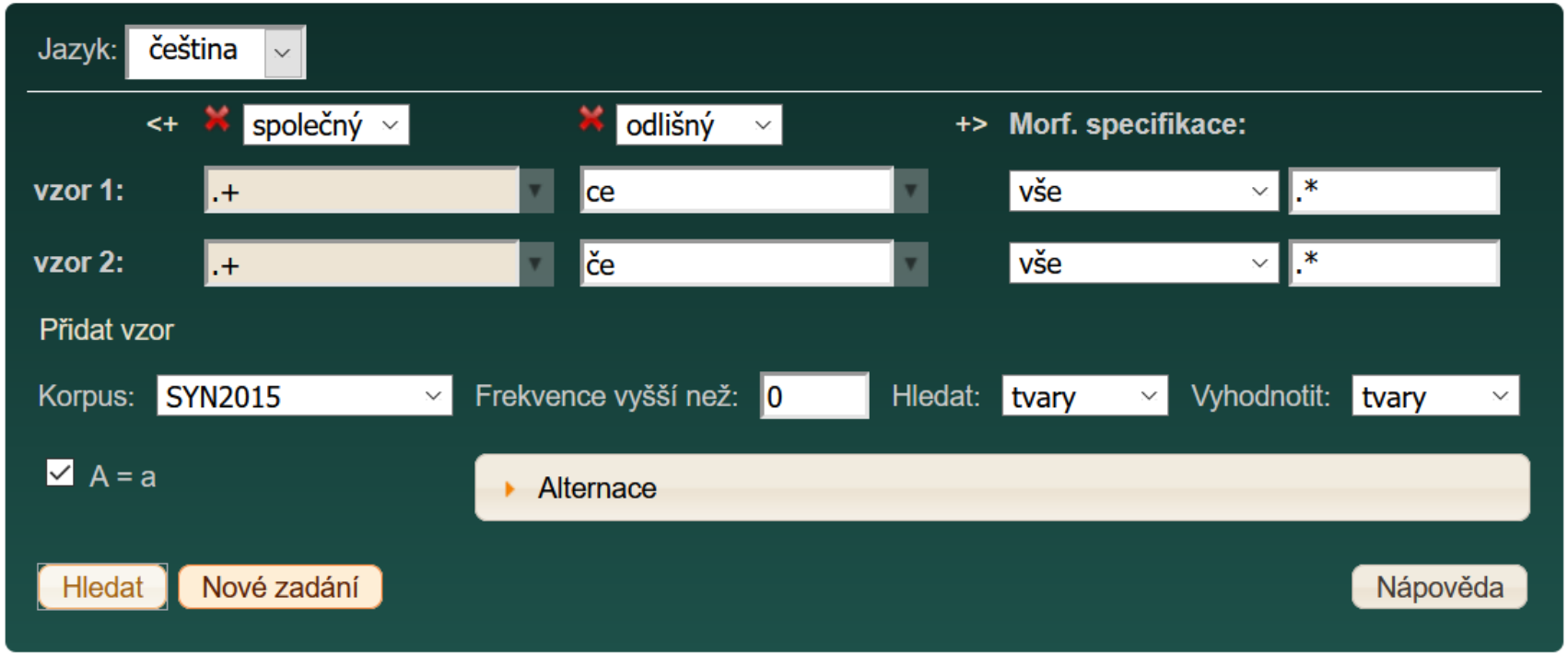
- Pozorujme spojitost mezi pravidlem o distribuci variantních koncovek vokativu singuláru maskulin životných a kodifikací.
- Pane *soud[cč]e*
- Který z tvarů je kodifikovaný?
- Umíme definovat důsledky systematické (nenahodilé – analogické) změny v kodifikaci?
- Lze někde v systému českého tvarosloví/derivace vidět působení tlaku analogie a v jejím důsledku i rozkolísání systému?

Otázky

- Je změna c/č v češtině pravidelná nebo nahodilá?
- Čím se řídí?
- Kde k ní dochází?

Hledáme dvojice slov, v nichž po c/č následuje přední vokál a jinak jsou identické

Morfio



Pozorování

- prá[cč]e
- ru[cč]e, kon[cč]e
- ot[cč]e, zástup[cč]e, chlap[cč]e, vůd[cč]e, …
- Závěr: Distribuce [cč] před předními vokály není řízena fonologicky, ale morfonologicky.
- Všimněme si, jak je tomu v případě [rř]:
- kobře, patře, Petře, doktore

Závěr

- Potřeba povědomí o možnostech a mezích formalizace (pravidla a výjimky, kontextová pravidla)
- Potřeba povědomí o rozdílech v terminologii (např. kmen, koncovka, …)
- Potřeba povědomí o fungování nástrojů (např. jednotlivé kroky automatické analýzy)
- Potřeba povědomí o technických zjednodušeních (např. jednoslovná morfologie)

Ke čtení

- OSOLSOBĚ, Klára a Karel PALA. Czech Stem Dictionary. Czech stem Dictionary. In *Sborník prací filozofické fakulty brněnské univerzity*. 1. vyd. Brno: Masarykova univerzita, Brno, 1993. s. 51-60, 10 s. ISBN 80-210-0883-0. ()
- OSOLSOBĚ, Klára, PALA, Karel, RYCHLÝ, Pavel. Frekvence vzorů českých sloves (na materiálu ČNK). *Slovo a slovesnost*, Praha: Akademie věd ČR, ÚJČ, 1998, roč. 98, č. 4, s. 265-277. ISSN 0037-7031.
- (http://sas.ujc.cas.cz/archiv.php?lang=en&art=3804)
- https://digilib.phil.muni.cz/bitstream/handle/11222.digilib/100316/A_Linguistica_46-1998-1_9.pdf?sequence=1
- Jaroslava Hlaváčová, Marie Mikulová, Barbora Štěpánková, Jan Hajič (2019): Modifications of the Czech morphological dictionary for consistent corpus annotation. In: *Jazykovedný časopis / Journal of Linguistics*, ISSN 0021-5597, vol. 70, no. 2, pp. 380-389