Přegenerovávání a podgenerovávání – dva problémy automatické analýzy přirozeného jazyka, konkrétně slovotvorby

PLIN033_3

PŘEGENEROVÁVÁNÍ

- *Formální definici (algoritmu) odpovídají jednotky, které tvoří homogenní skupinu (tu, kterou se prostřednictvím formálního zadání snažíme definovat), ale i jednotky, které jsou vůči této skupině heterogenní. Tento jev spadá na vrub obecné vlastnosti přirozeného jazyka, jíž je víceznačnost (homonymie) na všech úrovních.*

PODGENEROVÁVÁNÍ

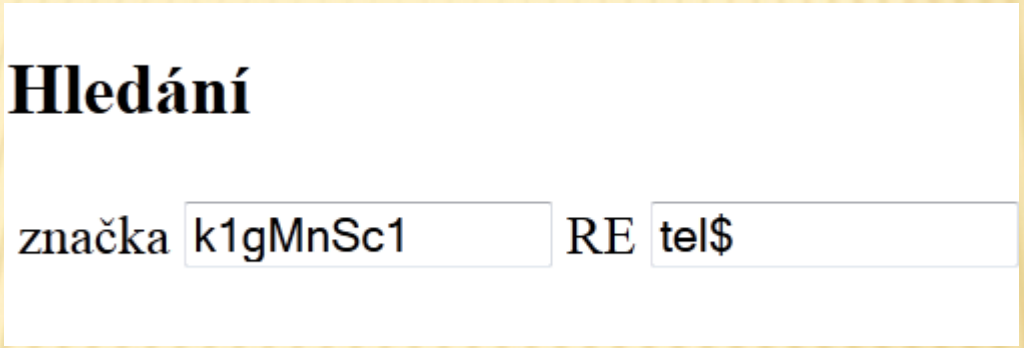
- *Rubem téže mince je tzv. podgenerovávání, tedy případ, kdy formální zadání je vymezeno příliš úzce, takže nejsou zachyceny jednotky, které se jeho prostřednictvím snažíme definovat.*

PŘÍKLADY PŘEGENEROVÁVÁNÍ Z MINULÝCH CVIČENÍ

- *Náboženství, nádeničení, …*
- *Klíč, míč, …*

POMOCÍ NÁSTROJE *DERIV* A *MORFIO* VYHLEDEJTE KANDIDÁTY NA ČINITELSKÁ JMÉNA NA *-TEL*

- Maskulina životná s koncovým řetězcem *tel*



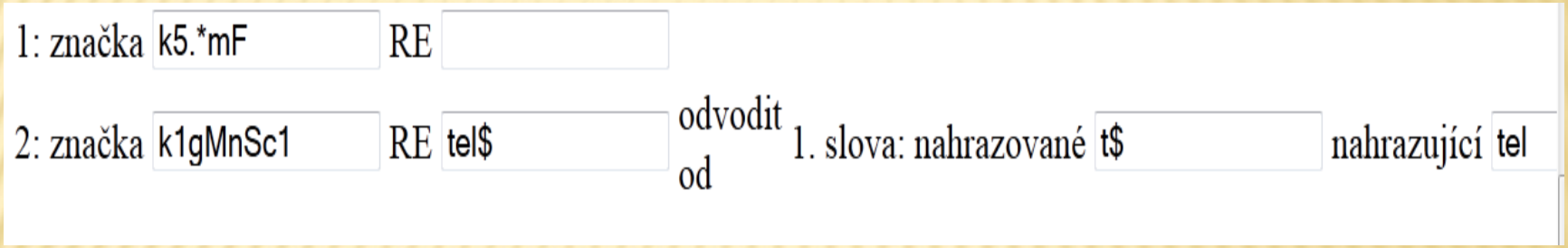
DERIV

- seznam

1		badatel
2		bavitel
3		bičovatel
4		bořitel
5		bouratel
6		branitel
7		brojitel
8		brzditel
9		buditel
10		budovatel
11		burcovatel
12		bydlitel
13		cenitel
14		cestovatel

DERIV
HLEDÁNÍ DVOJIC

- t$/k5.*mF>tel/k1gMnSc1



DERIV

- seznam

1		badat, badatel
2		bavit, bavitel
3		bičovat, bičovatel
4		bořit, bořitel
5		bourat, bouratel
6		brojit, brojitel
7		brzdit, brzditel
8		budit, buditel
9		budovat, budovatel
10		burcovat, burcovatel
11		bydlit, bydlitel
12		cenit, cenitel
13		cestovat, cestovatel
14		cvičit, cvičitel
15		čekat, čekatel
16		činit, činitel
17		čistit, čistitel

MORFIO

- Seznam

	vzor 1 (fq)	vzor 2 (fq)
1	živit (3800)	živitel (142)
2	zvěstovat (250)	zvěstovatel (49)
3	zůstavit (38)	zůstavitel (295)
4	zřizovat (555)	zřizovatel (599)
5	zprostředkovat (910)	zprostředkovatel (605)
6	zpracovat (2947)	zpracovatel (460)
7	zplodit (485)	zploditel (11)
8	znečišťovat (143)	znečišťovatel (147)
9	zmocnit (2601)	zmocnitel (29)
10	zhotovit (889)	zhotovitel (776)
11	zásobit (204)	zásobitel (15)
12	zasílat (719)	zasílatel (51)
13	zapisovat (1400)	zapisovatel (85)
14	zajistit (12412)	zajistitel (17)
15	vyzývat (1610)	vyzývatel (51)
16	vyšetřovat (2548)	vyšetřovatel (2394)
17	vystavovat (2286)	vystavovatel (477)
18	vysílat (3721)	vysílatel (40)
19	vypisovat (351)	vypisovatel (16)
20	vykořisťovat (86)	vykořisťovatel (59)
21	vychovat (1091)	vychovatel (389)
22	vyhlašovat (834)	vyhlašovatel (107)
23	vydražit (227)	vydražitel (34)
24	volit (5757)	volitel (117)
25	vnímat (10105)	vnímatel (43)
26	věznit (520)	věznitel (148)
27	věřit (29642)	věřitel (2047)

PŘEGENEROVÁVÁNÍ

- Přít/přítel

341		přisvědčovat, přisvědčovatel
342		přít, nepřítel
343		přít, přítel
344		přitakat, přitakatel

58	ručit (658)	ručitel (197)
59	rozhodovat (7902)	rozhodovatel (79)
60	přít (998)	přítel (23968)
61	přihlašovat (131)	přihlašovatel (30)
62	představit (22703)	představitel (6912)

DŮVODY PŘEGENEROVÁVÁNÍ

- Příliš široké formální vymezení
- Nemožnost užšího formálního vymezení

PODGENEROVÁVÁNÍ

- Kde jsou slova jako *ředitel, uchvatitel, šiřitel, majitel, pisatel, … ?*
- Zahrnutí alternací do vyhledávání jakožto prostředek zúžení definice hledaných jednotek.

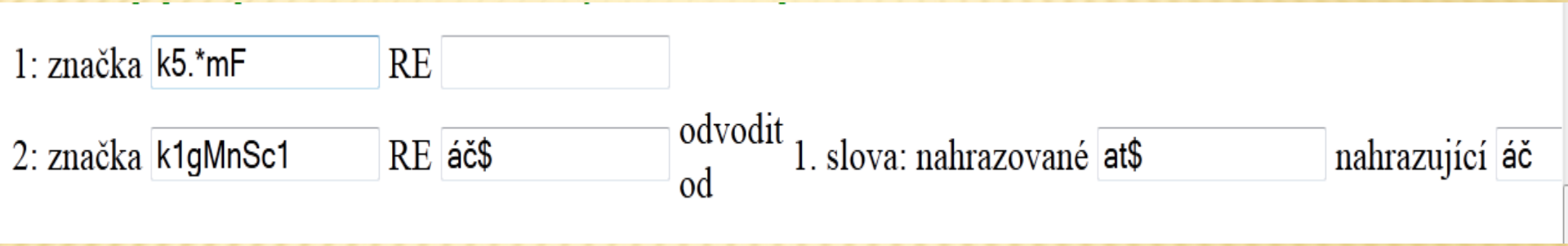
DERIVAČNÍ PRAVIDLA A VÝSLEDKY PRO DERIVACI SLOVESO – DĚJOVÉ JMÉNO NA *-TEL*

-tel

substituční pravidlo	příklad		dvojice	přegenerovávání
(?<!í)t$/k5.*mF >tel/k1gMnSc1[1]	*učit/učitel*	716	716	0
et$/k5.*mF>itel/k1gMnSc1	*velet/velitel*	20	16	4[2]
([ěí])t$/k5.*mF>itel/k1gMnSc1	*trpět/trpitel*	27	22	5[3]
át$/k5.*mF>atel/k1gMnSc1	*znát/znatel*	2	1	1[4]
át$/k5.*mF>ítel/k1gMnSc1	*přát/přítel*	2[5]	2	0
á(.)[iě]t$/k5.*mF>a$1itel/k1gMnSc1	*uchvátit/uchvatitel*	19	18	1[6]

VYHLEDÁVÁNÍ DVOJIC

- at$/k5.*mF>áč/k1gMnSc1



PŘEGENEROVÁVÁNÍ

Práce s obsahem souboru PLIN033/at_áč

Soubor **PLIN033/at_áč** byl úspěšně uložen (2 s).

Poznámkou je jednotlivý znak, jeden řádek může mít více poznámek.
Jako poznámku nelze použít dvojtečku, čárku a mezeru (budou-li zadány, budou ignorovány).

1		belhat, belháč
2		brnkat, brnkáč
3		brumlat, brumláč
4		bukat, bukáč
5		česat, česáč
6		hafat, hafáč
7		hmatat, hmatáč
8		chmatat, chmatáč
9		klepetat, klepetáč
10		kokrhat, kokrháč
11		kopat, kopáč
12		kovat, kováč
13		krkat, krkáč
14		orat, oráč

KLEPETÁČ

- Slovník

ssjc Slovník spisovného jazyka českého

klepetáč

-e m. expr. *velký rak:* chytil k-e za hlavu, aby nestřihl (Pujm.); --> expr. zdrob. **klepetáček,** -čka m. (mn. 1. -čci, -čkové 6. -čcích)

KRKÁČ

- slovník

ssjc Slovník spisovného jazyka českého

***krkáč**

-e m. expr. *lakomec, chamtivec, krkoun* (Herrm., Šlej.)

DŮVODY

- Polyfunkčnost prostředku (-á-č x –áč)
- Závisí na mimojazykových znalostech
- Obtížně se formálně definuje

PODGENEROVÁVÁNÍ

- Nedostatky ve formální definici
- Nepravidelnosti (*vozač, trubač*)
- Jednotky nejsou zachyceny ve slovníku
- Jednotkám nezachyceným ve slovníku chybí interpretace na úrovni lemmatu a morfologické značky

MORFIO

- *kout/kouč, klít/klíč, sálat/salač*
- Propria: *máchat/Machač, tykat/Tykač, dědit/Dědič, pískat/Piskač, kopat/Kopač, klapat/Klapač, kovat/Kovač, pleskat/Pleskač, bílit/Bilič*

TYPY PŘEGENEROVÁVÁNÍ

- hláskové alternace kořenového vokálu u derivátů od sloves III. třídy podle kmene prézentního (vzor *krýt*)
- hláskové alternace kořenového vokálu u ostatních tříd a vzorů
- hláskové alternace kmenotvorného vokálu u ostatních tříd a vzorů

ALTERNACE KOV U DERIVÁTŮ SLOVES PODLE *KRÝT*

- *hrát/hráč*
- *chcát/chcáč*
- *? pít/píč*
- *? pět/pěč*
- *? sít/síč*

V KORPUSECH LZE NAJÍT (SYN)

- *pít* (čaj)/ *čajpíč*
- *žít/žíč*
- ***!*** *šít/šič*

" Po kafi mám blijavku , soudruzi ! Máte čaj . . . ? Já jsem . . . ***čajpíč*** ! Kdo je tady eště . . . čaj . . . ? "

Reflex, č. 45/2002	, že spadáme do kategorie	žíčů	/žíčů/X@----------	. Jako je někdo optimista
Reflex, č. 45/2002	flegmatik , tak my jsme	žíči	/žíči/X@----------	. A když už se
Reflex, č. 45/2002	už se jednou člověk takovým	žíčem	/žíčem/X@----------	narodí , tak jím nikdy

Blesk, 17. 1. 2000	měsíc . " Rekvalifikaci na	šiče	/šiče/X@----------	mi nabídli , když jsem
Blesk, 17. 1. 2000	Loni bylo na rekvalifikační kurs	šiče	/šiče/X@----------	, který zajišťuje zmíněná firma
Blesk, 18. 5. 2000	zájem o práci šiček a	šičů	/šičů/X@----------	, kterých tady pracuje sto

A KROMĚ TOHO U NEŽIVOTNÝCH MÁME

- *bít/bič*
- *rýt/rýč*

VŠIMNĚME SI DVOJIC

- *vypr**á**vět/vypr**a**věč |vypr**á**věč*
- *vyjedn**á**vat/vyjedn**a**vač |vyjedn**á**vač*

IJP

- http://prirucka.ujc.cas.cz/?id=730#nadpis14
- **2 Střídání krátkých a dlouhých samohlásek při tvoření slov**
- Příklady nikoli pravidla (?seznamy výjimek)

LITERATURA

- OSOLSOBĚ, Klára. *Morfologie českého slovesa a tvoření deverbativ jako problém strojové analýzy češtiny*. 1. vyd. Brno: Masarykova univerzita, 2011. 220 s. Spisy FF MU v Brně č. 401. ISBN 978-80-210-5565-0.
- CVRČEK, Václav: *Co je nového v ČNK II.* KORPUS – GRAMATIKA – AXIOLOGIE 7/ 2013, 95-97.

ÚKOL NA 30.10. 2013

- Pomocí nástrojů *Deriv* a *morfio* vyhledejte kandidáty na trojice sloveso-činitelské jméno na –č – ženský protějšek na –čka (sloveso – {jméno prostředku na –č} – jméno prostředku na -čka).
- Popište případy přegenerování popř. podgenerování