PLIN033_3

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

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*

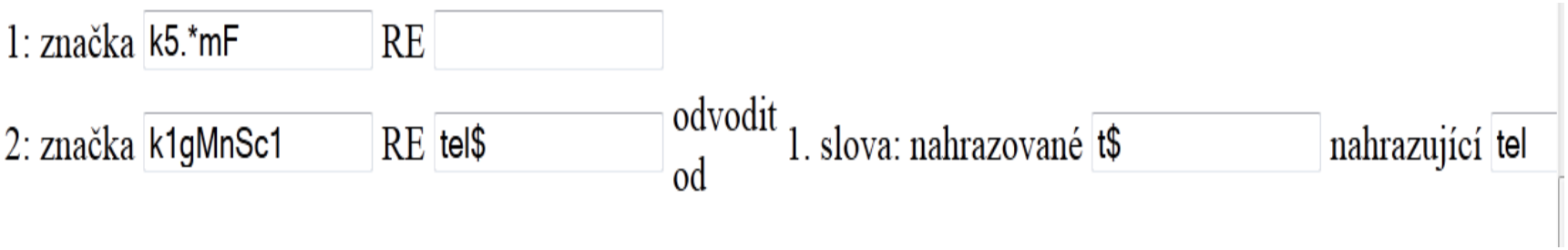
Deriv

- 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]

Derivační pravidla a výsledky pro derivaci sloveso – dějové jméno na *-tel*

í(.)it$/k5.*mF> i$1itel/k1gMnSc1	*šířit/šiřitel*	7	7	0
í(.)it$/k5.*mF>e$1itel/k1gMnSc1	*řídit/ředitel*	2	2	0
ou(.)[ie]t$/k5.*mF>u$1itel/k1gMnSc1	*vykoupit/vykupitel*	11	9	2[167]
á(.|ch)at$/k5.*mF>a$1atel/k1gMnS1	*dodávat/dodavatel*	57	57	0
í(.)at$/k5.*mF>i$1atel/k1gMnSc1	*užívat/uživatel*	15	15	0
í(.)at$/k5.*mF>ě$1atel/k1gMnSc1	*přispívat/přispěvatel*	4	4	0
ý(.)at$/k5.*mF>y$1atel/k1gMnSc1	*obývat/obyvatel*	7	7	0
é(.)at$/k5.*mF>e$1atel/k1gMnSc1	*obléhat/oblehatel*	2	1	1[168]
ou(.)at$/k5.*mF>u$1atel/k1gMnSc1	*zkoumat/zkumatel*	1	1	0

Derivační pravidla a výsledky pro derivaci sloveso – dějové jméno na *-tel*

jmout$/k5.*mF>jatel/k1gMnSc1[169]	*pronajmout/pronajatel*	1	1	0
ýt$/k5.*mF>ytel/k1gMnSc1	*vydobýt/vydobytel*	1	1	0
ít$/k5.*mF>ajitel/k1gMnSc1	*mít/majitel*	1	1	0
ést$/k5.*mF>editel/k1gMnSc1	*provést/proveditel*	1	1	0
ct$/k5.*mF>žitel/k1gMnSc1	*přemoct/přemožitel*	1	1	0
^psát$/k5.*mF> pisatel/k1gMnSc1	*psát/pisatel*	1	1	0
CELKEM	**22 pravidel**	**898**	**884**	**14**
Výjimky	*neumětel, postihatel, dosažitel, ...*			

Přegenerované doklady

- *mučet/mučitel, proset/prositel, těžet/těžitel, zcizet/zcizitel.*
- *mocnět/mocnitel, pět/pitel, zmocnět/zmocnitel, pro-sít/prositel, učít/učitel.*
- *dát/datel.* Tento případ přegenerovávání by bylo možno eliminovat aplikací podmínky, že substantivum musí být skloňováno podle vzoru *muž*, již by bylo možno zadat v případě, že bychom pracovali se strojovým slovníkem značkovaným tak, že by součástí značky byla i informace o flektivním typu (vzor).
- Jedná se o substantiva *přítel* a *nepřítel*. V praxi automatické morfologické analýzy (lemmatizace) nepanuje jednota v interpretaci derivátů se záporkou *ne-* (srv. Osolsobě 2007[1]). Řešení tohoto problému přesahuje záměr této práce.
- *zařádit/zařaditel.*
- *boudit/buditel, moučit/mučitel.*
- *ohlédat/ohledatel.*

Přehled alternací

alternace KmV	příklad
0-i	*přemoc0t-přemožitel*
á-a	*psát-pisatel*
e-i/ě-i	*velet-velitel/ trpět-trpitel*
í-i	*křtít-křtitel*
mou-a	*pronajmout-pronajatel*
alternace KoV	
á-a	*uchvátit-uchvatitel*
é-e	*obléhat-oblehatel*
í-i	*šířit-šiřitel*
í-e	*řídit-ředitel*
í-ě	*přispívat-přispěvatel*
ou-u	*vykoupit-vykupitel*
ý-y	*hýbat-hybatel*
í-aj	*mít-majitele*
0-i	*p0sát-pisatel*
á-í	*přát-přítel*

Přehled alternací

alternace kf	
c-ž	*přemoc0t-přemožitel*
h-ž	*dosáhnout-dosažitel*[158]
konekt	
nou-a	*postihnout-postihatel*
nou-i	*dosáhnout-dosažitel*
alternace na švu prefix/kořen (P0e, ki)[159]	

Vyhledávání dvojic

- **at$/k5.*mF>áč/k1gMnSc1**

1: značka k5.*mF RE

2: značka k1gMnSc1 RE áč$ odvodit od 1. slova: nahrazované at$ nahrazující áč

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)

Úkol na 29.10. 2014

- Pomocí nástrojů *Deriv* a *morfio* vyhledejte kandidáty na dvojice sloveso-jméno prostředku na –dlo.
- Popište případy přegenerování popř. podgenerování