Využití corpus driven metod při corpus based výzkumu
Přegenerovávání a podgenerovávání: Skylla a Charybda automatické morfologické analýzy

Klára Osolsobě

Ústav českého jazyka FF MU

Abstract

Overgeneration is a property of formal rules which does not cover the exact language data it was designed for. It is equivalent to low precision and occurs when a formal rule (corpus query) is too widely defined. Undergeneration is equivalent to low recall and occurs when a formal rule (corpus query) is too narrowly specified. Both are caused by the ambiguity of natural language. In this article we shall demonstrate how to use corpus driven method in optimization of retrieval technique for corpus based analysis. On a specific example of retrieval of candidates for a word formation model (*kutil*) we shall show how to use observation of corpus data for progressive specification of corpus query.

Klíčová slova: *korpus, corpus based, corpus driven, přegenerovávání, podgenerovávání, lemma, tag, slovotvorba*.

Keywords: *Corpus, corpus based, corpus driven, overgeneration, undegeneration, lemma, tag, word formation*.

Úvod

Jedním z podstatných rysů aplikací automatické analýzy přirozeného jazyka je tzv. přegerovávání. Formální definici 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 nejednoznačnost (homonymie) na všech úrovních. 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.

Na konkrétním příkladu ukážeme postup optimalizace vyhledávání dat pro korpusově založený (corpus based) výzkum slovotvorby, který vychází z korpusově řízené (corpus driven) metody zpřesňování formálního zadání na základě pozorování přegenerovaných dat.

Korpusově řízená metoda optimalizace postupů vyhledávání dat v jazykových korpusech[1]

Při vyhledávání jazykových dat používáme při práci s jazykovými korpusy počítačové programy (korpusové manažery), které umožňují klást dotazy, jimiž se snažíme co možná nejpřesněji definovat vlastnosti (formální) jednotek jazyka, které chceme v korpusu vyhledat.
V tomto příspěvku se snažíme ukázat, jak lze postupným upřesněním formálního zadání dojít k lepším datům, která jsou podkladem pro lingvistický výzkum.
Chceme-li minimalizovat přegenerovávání, pak je třeba formální definici pokud možno zúžit.
Naopak chceme-li minimalizovat podgenerovávání, musíme mít na zřeteli, že příliš úzké zadání může mít za následek ztrátu dat.
Ukažme si, jak funguje pře- a pod-generování na praktickém a celkem jednoduchém příkladu deverbativ tvořených příponou *-tel*. Formální jednoduchá definice zní, že kandidáty lze najít, vyhledáme-li všechna maskulina životná zakončená řetězcem ***.*tel***. Při použití takové definice budou případy přegenerovávání slova jako *datel* nebo proprium *Obrtel*. Přegenerování se nevyhneme ani tehdy, zúžíme-li definici na případy, kdy před ***tel*** předchází ***[ai]***. Zůstane přegenerovaný ***datel,*** a navíc podgenerovaný ***přítel*** a ***neumětel.***
Nyní si ukážeme složitější případ. Budeme používat formální zápis v jazyce **cql** (**c**orpus **q**uery **l**anguage).[2]

Postup optimalizace vyhledávání substantiv typu *kutil*

Našim cílem je vyhledat v korpusu kandidáty na **jména osob odvozená od l-ových příčestí (konvertovaná)**. Máme-li označkovaný korpus, pak se můžeme při zadání dotazu opřít a) o morfologickou značku a b) o lemma. Chceme tedy vyhledat maskulina životná, jejichž základní tvar (lemma) končí na *l*.

[1] Praktické ukázky jsou vybrány ze synchronního korpusu psané češtiny (SYN2010) – http://ucnk.ff.cuni.cz/. Při práci s korpusem v ukázkách (obrázky) používáme verzi korpusového manažeru přístupného z webového rozhraní NoSketch Engine na adrese http://korpus.cz/corpora/.
[2] Bližší informace k jazyku CQL i k použitému morfologickému značkování lze nalézt zde: http://ucnk.ff.cuni.cz/bonito/regular.php a http://ucnk.ff.cuni.cz/bonito/znacky.php.

Dotaz v CQL bude vypadat následovně: **[lemma=".*l" & tag="NNM.*"]** a po jeho zadání do dotazovacího řádku korpusového manažeru získáme konkordanci všech výskytů substantiv, maskulin životných, jejichž základní tvar končí na *l* (obrázek 1).

Obrázek 1

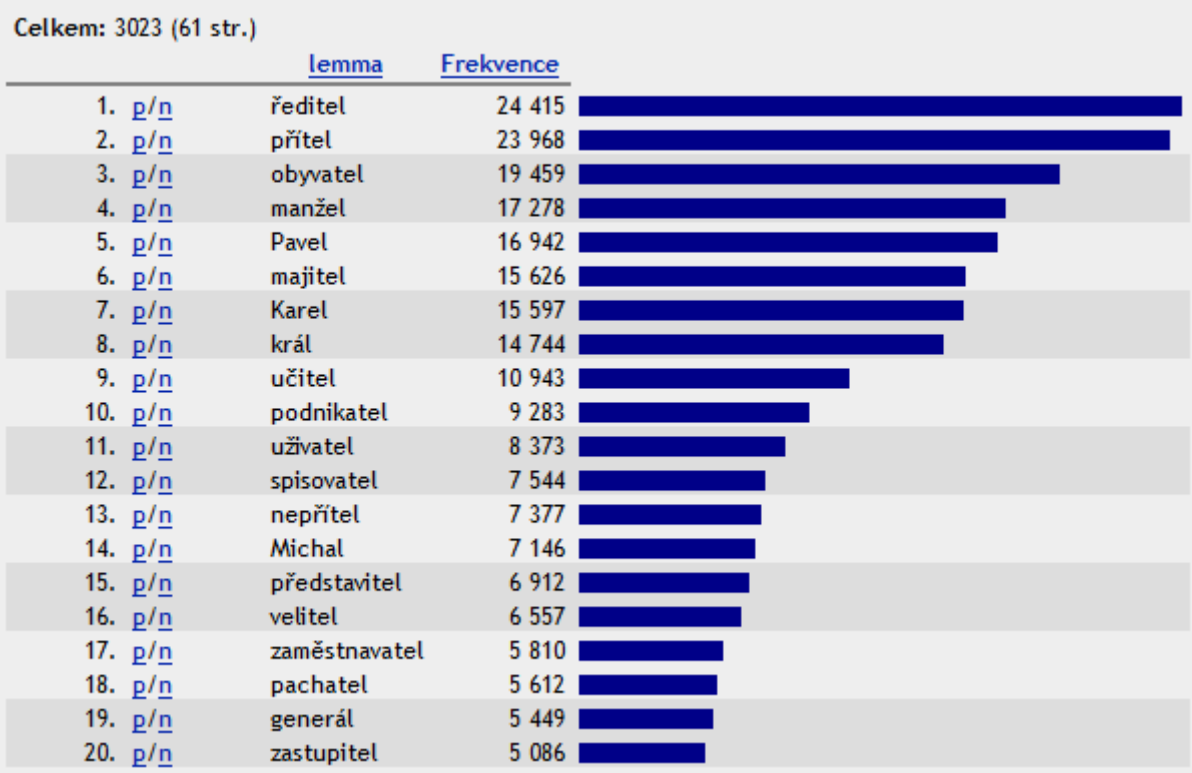
Příběhy Sherlocka Holmese	by asi udivily i jeho	učitele	/učitel/NNMP4-----A-----	. " " Neptal ses
Příběhy Sherlocka Holmese	se mohl s tím tvým	přítelem	/přítel/NNMS7-----A-----	seznámit ? " " Určitě
Příběhy Sherlocka Holmese	Je schopen podat svému nejlepšímu	příteli	/přítel/NNMS3-----A-----	nejnovější rostlinný alkaloid , ne
Příběhy Sherlocka Holmese	přišoupl nohou . " Můj	přítel	/přítel/NNMS1-----A-----	hledá byt , a poněvadž
Příběhy Sherlocka Holmese	vzbudit zájem i toho nejpovrchnějšího	pozorovatele	/pozorovatel/NNMS4-----A-----	. Byl šest stop vysoký
Příběhy Sherlocka Holmese	krásného počasí , neměl jsem	přátel	/přítel/NNMP2-----A-----	, kteří by svými návštěvami
Příběhy Sherlocka Holmese	spolubydlící nemá tak jako já	přátel	/přítel/NNMP2-----A-----	. Potom jsem však zjistil
Příběhy Sherlocka Holmese	aby si z nich schopný	pozorovatel	/pozorovatel/NNMS1-----A-----	neudělal přesný obraz . "
Příběhy Sherlocka Holmese	mlčení , co si jeho	přítel	/přítel/NNMS1-----A-----	myslí , to sice vypadá
Příběhy Sherlocka Holmese	. " Lecoq byl ubohoučký	packal	/packal/NNMS1-----A-----	, " řekl zlostně .
Příběhy Sherlocka Holmese	pokoje , a podával mému	příteli	/přítel/NNMS3-----A-----	dopis . To byla příležitost
Příběhy Sherlocka Holmese	Mohu se vás zeptat ,	příteli	/přítel/NNMS3-----A-----	, " řekl jsem líbezně
Příběhy Sherlocka Holmese	Yardu , " poznamenal můj	přítel	/přítel/NNMS1-----A-----	, " on a Lestrade
Příběhy Sherlocka Holmese	on a Lestrade jsou jednookými	králi	/král/NNMP7-----A-----	mezi slepými . Oba jsou
Příběhy Sherlocka Holmese	. " " Můj drahý	příteli	/přítel/NNMS3-----A-----	, co z toho mám
Příběhy Sherlocka Holmese	tohohle ! " odpověděl můj	přítel	/přítel/NNMS1-----A-----	a ukázal na pěšinu .
Příběhy Sherlocka Holmese	na pěšinu . " Stádo	buvolů	/buvol/NNMP2-----A-----	by to tu nemohlo víc
Příběhy Sherlocka Holmese	detektiv . Tvářil se jako	majitel	/majitel/NNMS1-----A-----	panoptika , který se vychloubá
Příběhy Sherlocka Holmese	vám , " odpověděl můj	přítel	/přítel/NNMS1-----A-----	. " Počínáte si tak
Příběhy Sherlocka Holmese	zahradní pěšině šli jako dva	přátelé	/přítel/NNMP1-----A-----	- pravděpodobně byli do sebe

Je patrné, že vidíme téměř samé přegenerované doklady (výjimkou je substantivum *packal* na 10. řádku konkordančního seznamu na obrázku 1). Zobrazíme-li si frekvenční distribuci lemmat (obrázek 2), pak zjistíme, že pro získání nějakých relevantních výsledků by bylo třeba projít seznam 3023 lemmat ručně. (Až na 102. řádku frekvenčního seznamu lemmat objevíme příjmení *Navrátil*, které splňuje podmínky zadání a zároveň jde o případ, který jsme se snažili prostřednictvím zadání vyhledat.[3]) Takový postup je a) velmi pracný a b) jako každá ruční analýza lze předpokládat, že dojde k chybám zaviněným „lidským faktorem“ (nepozornost). Z tohoto důvodu se pokusíme postup vyhledávání optimalizovat.

Z frekvenční distribuce lemmat plyne, že zadání vede k masivnímu přegenerování. Dotaz nelze zpřesnit tím, že bychom zadali omezení pouze na substantiva skloňovaná podle vzoru *pán*, neboť vzor není součást morfologické značky. Vidíme totiž, že přegenerovávání spadá na vrub zejména této chybě (slova jako ***ředitel, přítel, obyvatel*** se skloňují podle vzoru *muž*), ale nejen jí (slova jako ***Karel, Pavel, Michael, generál*** nejsou hledanými doklady, ačkoliv se skloňují podle vzoru *pán*).

[3] Na 84. řádku se objevuje příjmení *Doležal*, které je sice odvozeno od slovesa, ale nelze je synchronně bez problémů formálně interpretovat jako konvertované l-ové příčestí.

Obrázek 2

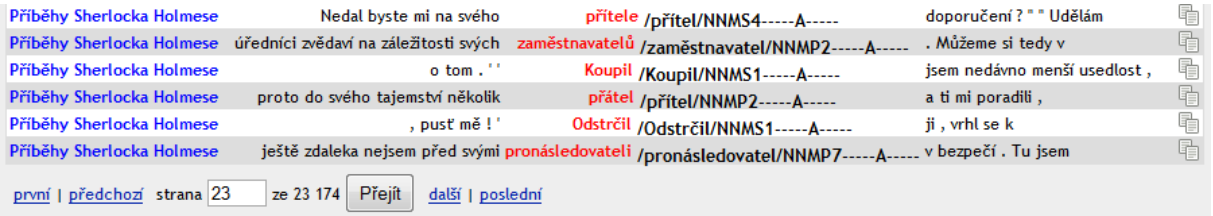


Celkem: 3023 (61 str.)

		lemma	Frekvence
1.	p/n	ředitel	24 415
2.	p/n	přítel	23 968
3.	p/n	obyvatel	19 459
4.	p/n	manžel	17 278
5.	p/n	Pavel	16 942
6.	p/n	majitel	15 626
7.	p/n	Karel	15 597
8.	p/n	král	14 744
9.	p/n	učitel	10 943
10.	p/n	podnikatel	9 283
11.	p/n	uživatel	8 373
12.	p/n	spisovatel	7 544
13.	p/n	nepřítel	7 377
14.	p/n	Michal	7 146
15.	p/n	představitel	6 912
16.	p/n	velitel	6 557
17.	p/n	zaměstnavatel	5 810
18.	p/n	pachatel	5 612
19.	p/n	generál	5 449
20.	p/n	zastupitel	5 086

Při procházení konkordancí si můžeme (na stránce 23 konkordančního seznamu) všimnout i zdánlivě správných dokladů (obrázek 3). Při bližším pozorování příslušného kontextu ovšem zjistíme, že jde o chyby v disambiguaci (viz interpretace tvarů *koupil* a *odstrčil* na 3. a 5. řádku konkordančního seznamu na obrázku 3).

Obrázek 3



Příběhy Sherlocka Holmese	Nedal byste mi na svého	přítele	/přítel/NNMS4-----A-----	doporučení ? " " Udělám
Příběhy Sherlocka Holmese	úředníci zvědaví na záležitosti svých	zaměstnavatelů	/zaměstnavatel/NNMP2-----A-----	. Můžeme si tedy v
Příběhy Sherlocka Holmese	o tom . ' '	Koupil	/Koupil/NNMS1-----A-----	jsem nedávno menší usedlost ,
Příběhy Sherlocka Holmese	proto do svého tajemství několik	přátel	/přítel/NNMP2-----A-----	a ti mi poradili ,
Příběhy Sherlocka Holmese	, pusť mě ! '	Odstrčil	/Odstrčil/NNMS1-----A-----	ji , vrhl se k
Příběhy Sherlocka Holmese	ještě zdaleka nejsem před svými	pronásledovateli	/pronásledovatel/NNMP7-----A-----	v bezpečí . Tu jsem

první | předchozí strana 23 ze 23 174 Přejít další | poslední

Shrňme tedy, že s ohledem na skutečnost, že není možné jednoduše pomocí morfologické značky vyhledávat podle vzoru, dochází při zvoleném zadání dotazu k masivnímu přegenerovávání. Dále je vidět, že přegenerovává i disambiguace.

Je tedy žádoucí hledat jiný (lepší) postup. Optimalizace může vycházet z pozorování přegenerovaných dat (corpus driven) a následné formulace pravidel transformovatelných do

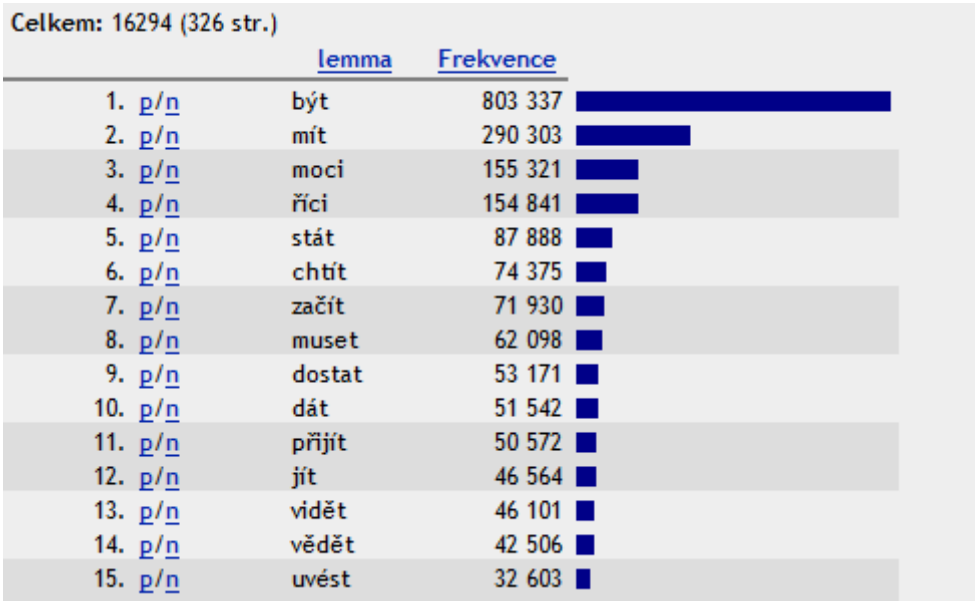
optimalizovaných formálních definic pro hledání kandidátů reprezentujících příslušný slovotvorný typ.
Pozorujeme, že většina přegenerovaných případů jsou deverbativa tvořená sufixem *-tel*. Položíme si tedy otázku, zda mohou l-ová příčestí v češtině končit např. na řetězec *tel* (ale i např. *rel, vel, ael,* neboť přegenerovaná jsou i lemmata vlastních jmen jak *Karel, Pavel, Michael*). Odpověď lze následovně zahrnout do dotazu v cql jako formální podmínku pro vyhledávání.

Popis postupu pro získání odpovědi na základě pozorování korpusových dat

Na otázku, mohou-li l-ová příčestí v češtině končit např. na řetězce *tel, rel, vel, ael,* lze odpovědět takto: a) mohou a v korpusu jsou doložena, b) mohou, ale v korpusu nejsou doložena, c) nemohou a tudíž v korpusu nejsou doložena.
Hledáme ospravedlnění kladné odpovědi na otázku a). Slovní formulace dotazu pro získání dat z korpusů bude, že hledáme l-ová příčestí taková, že před řetězcem *l*, popřípadě *la, lo, li, ly* (*l*+rodová koncovka) předchází řetězec *tel, rel, vel, ael*.
V cql bude dotaz vypadat takto: **[word="((.*[atrv]el)|(.*[atrv]el[aoiy]))" & tag="V[pq].*"]** a zadáme-li jej ve formě dotazu korpusovému manažeru, bude výsledek vyhledávání nulový (prázdný seznam konkordančních řádků).
Je-li výsledkem prázdný seznam, mělo by platit buď b), nebo c). Prohledáváním ještě větších korpusů by bylo možné pokračovat a snažit se dokázat, že platí za b). Jde o neefektivní postup s malou pravděpodobností úspěchu, neboť ze zkušenosti práce s korpusy je známo, že výjimky u frekventovaných tvarů bývají rovněž frekventované a l-ové příčestí je frekventovaný tvar. Můžeme ovšem hledat nějaké zobecnitelné formální vlastnosti l-ových příčestí a pokusit se tak dokázat, že platí c).
Viděli jsme, že relevantní výsledek jsme docílili, když jsme se zabývali otázkou, co předchází před tvarovou koncovkou *-l* v českých l-ových příčestích. Odpovědi mohou být např. tyto: A) libovolná samohláska/souhláska (grafém), B) pouze některé samohlásky/souhlásky (grafémy). Jednoduchou empirickou evidenci získáme z korpusů, a to tak, že se podíváme, jaká je situace u l-ových příčestí (obrázek 4).
Vidíme, že před tvarovou koncovkou *-l* může předcházet např. dlouhé *á*, krátké *[aeěiyu]*, že se vůbec nevyskytují dlouhé samohlásky (s výjimkou *á*), z krátkých se nevyskytuje *o* (kromě slovenského *bol*). Pokud bychom chtěli zjistit, které souhlásky se vyskytují v českých l-ových

příčestích před tvarovou koncovkou *-l*, pak by dotaz v cql vypadal následovně: [tag="Vp.*" & word!="((.*[aáeěiyu]l)|(.*[aáeěiyu]l[aoiy]))"].[4]

Obrázek 4



Celkem: 16294 (326 str.)

		lemma	Frekvence
1.	p/n	být	803 337
2.	p/n	mít	290 303
3.	p/n	moci	155 321
4.	p/n	říci	154 841
5.	p/n	stát	87 888
6.	p/n	chtít	74 375
7.	p/n	začít	71 930
8.	p/n	muset	62 098
9.	p/n	dostat	53 171
10.	p/n	dát	51 542
11.	p/n	přijít	50 572
12.	p/n	jít	46 564
13.	p/n	vidět	46 101
14.	p/n	vědět	42 506
15.	p/n	uvést	32 603

Nyní si „pozveme na pomoc“ systematický popis morfologie českých sloves. Kmenový vokál (KmV) pro tvary od kmene minulého konkrétně l-ových příčestí, mohou být pouze krátké *[aieě]* a *(n)u*, nikdy *o*. Dlouhé *á* mají pouze nepravidelná slovesa jako *bál (se)*, *stál* a z pravidelných jen *zdál se* a *(u)dál se*. Kromě toho ovšem taky případy, kdy *á* je kořenový vokál (KoV), tedy sloveso patří ke vzoru *krýt* III. tř. – *hrál, sál, vál, smál se* atd. Všechny ostatní KoV ve III. tř. u vzoru *krýt* jsou krátké (zkrácené oproti infinitivu), vyskytuje se pouze *[eěiuy]* (*klel, děl, žil, zul, myl*) a obdobná omezení nacházíme i u nepravidelných (atematických) sloves *byl, měl, chtěl* atd. Souhlásku mají slovesa I. třídy vzorů *nést, péct* a mohou ji mít (nemají-li kmenotvorné *-nu-*) slovesa II. třídy vzoru *tisknout* (*tiskl, nadchl se, zestárl*) a některá vzoru *začít* (*zapl, vypl,* i moravské *rožl*). S ohledem na masivní přegenerovávání substantiv na *tel* nás zajímá především KoV a KmV *[eě]*.

KmV *[eě]* mohou mít slovesa I. třídy podle kmene přítomného vzoru *umřít* (*umřel, vytřel, pomlel*), IV. třídy podle kmene přítomného vzorů *trpět, sázet* (*vrtěl, hleděl, probděl*).

KoV *[eě]* mohou mít slovesa III. třídy podle kmene přítomného vzoru *kryl* (*děl, pěl, klel, zasel*).

[4] Projdeme-li seznam tvarů, pak zjistíme, že omezení se bude týkat grafémů [*ďjlnňqřťw*]. Ve sledovaném korpuse nenajdeme ani doklady [*gx*], nicméně l-ová příčestí sloves *grgnout* a *exnout* lze najít např. na internetu.

Bližší analýza možného okolí *e/ě* v roli KmV a KoV

Vyhledáme v korpusu všechna l-ová příčestí končící na *el*. V cql bude dotaz vypadat takto: [word="((.*el)|(.*el[aoiy]))" & tag="V[pq].*"]. Část seznamu lemmat podle frekvence vidíme na obrázku 5.

Obrázek 5

Celkem: 5958 (120 str.)

		lemma	word	Frekvence
1.	p/n	muset	musel	19 986
2.	p/n	přijít	přišel	17 754
3.	p/n	jít	šel	11 055
4.	p/n	muset	museli	10 507
5.	p/n	muset	musela	10 095
6.	p/n	najít	našel	8 065
7.	p/n	myslet	myslel	7 036
8.	p/n	zemřít	zemřel	6 589
9.	p/n	odejít	odešel	6 256
10.	p/n	otevřít	otevřel	6 217
11.	p/n	slyšet	slyšel	6 093
12.	p/n	vyjít	vyšel	5 257
13.	p/n	držet	držel	5 084
14.	p/n	pomyslet	pomyslel	4 500
15.	p/n	přijet	přijel	4 400
16.	p/n	zmizet	zmizel	4 030
17.	p/n	otevřít	otevřela	3 851
18.	p/n	muset	muselo	3 824
19.	p/n	myslet	myslela	3 742
20.	p/n	jet	jel	3 737

Na základě pozorování lemmat můžeme formulovat pravidla distribuce *e/ě* v roli KmV a KoV.

Před KmV *-e-* stojí v češtině pouze *[cčjlřsšzž]*, před KoV *-e-* stojí v češtině pouze *[lsz]*. Stojí-li před hláskou *e* grafémy *[pbvtdnm]*[5], pak se vždy graficky realizuje jako *ě*.

Na základě empirie můžeme tedy tvrdit, že aby mohlo být substantivum konvertovaným l-ovým příčestím, pak musí splňovat tyto podmínky: před *l* může ze samohlásek předcházet pouze *[aáeěiuy]*; před *l* nemůže předcházet *[bpvmdtnkghr]e*; před *l* nemůže předcházet *[aáeěéiíoóuůyý][aáeěiuy]*.

Formulace dotazu s využitím formálních vlastností pozorovaných dat

[5] V seznamu neuvádíme grafém *f*, protože na základě analýzy dat z korpusu SYN se nám nepodařilo najít žádný doklad l-ového příčestí slovesa s l-ovým příčestím na *.*fěl* a v korpusu czTenTen12 jsme nalezli pouze překlepy. (Dotaz v cql by vypadal takto: [word="((.*fěl)|(.*fěl[aoiy]))"].)

Do dotazu v jazyce cql zahrneme podmínky pro vyhledávání substantiv maskulin životných, která mají mít výše uvedené formální vlastnosti (lemma končí na *l* a zároveň před *l* předchází pouze některé přesněji definované kombinace grafémů). S ohledem na fakt, že v korpusech se poměrně často vyskytují chyby v desambiguaci homonymních tvarů l-ové příčestí/ tvar konvertovaného jména, omezíme vyhledávání jen na nehomonymní tvary. Přehled homonymních tvarů uvedeme v tabulce 1 (tučně a kurzívou). Jsme si vědomi také toho, že pomineme deriváty od sloves, u kterých se tvarová koncovka *-l* připíná bezprostředně ke kořenové souhláskové finále (slovesa I. třídy slovesné podle vzorů *nést, péci* a některá slovesa II. třídy slovesné podle vzorů *tisknout* a *začít*). Od těchto sloves se totiž jména osob sice okrajově tvořit mohou, tvoří se ovšem konverzí substandardního tvaru l-ového příčestí bez tvarové koncovky *-l*, jak dosvědčují např. apelativa *vyklouz, zběh* a snad i kompozita *břichopas, mrakotřas* nebo propria *Proklouz, Skoněspad, Vozembouch* atd. (více Osolsobě 2011: 61n).

V cql bude dotaz vypadat takto: **[word=".*[aáeěiuy]l((ové)|(ů)|(ům)|(ech)|(e)|(ovi))" & word!=".*[bpfvmdtnkghr]el((ové)|(ů)|(ům)|(ech)|(e)|(ovi))" & word!=".*[aáeěéiíoóuůyý][aáeěiuy]l((ové)|(ů)|(ům)|(ech)|(e)|(ovi))" & tag="NNM.*"]**

Tabulka 1

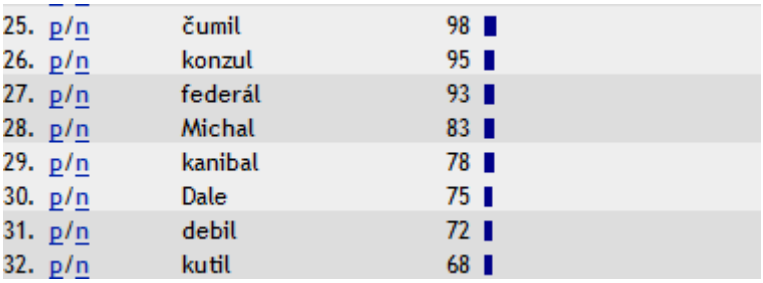
N. A.	***kutil***	kutilové
G.	***kutila***	kutilů
D.	kutilu kutilovi	kutilům
V.	kutile	kutilové
L.	kutilu kutilovi	kutilech
I.	kutilem	***kutily***

Na obrázku 6 vidíme, že se již na první stránce konkordancí objevuje alespoň jeden relevantní doklad v podobě tvaru *hýřilové*.

Obrázek 6

Při zobrazení frekvenční distribuce lemmat zjistíme, že stále je dost případů, které jsme nehledali (přegenerovávání), nicméně máme už i správné výsledky jako *čumil* a *kutil* (obrázek 7).

Obrázek 7



Ruční analýzou dat (631 lemmat) získáme jak apelativa, tak především propria (mezi nimi je hledaný slovotvorný model bohatě zastoupen). Nacházíme lemmata jako *kutil, čumil, Nezval, Hrabal, Pospíšil, čmuchal, ožrala, břídil, (Doležal), Musil, Prášil, patolízal, šťoural* a další (řazeno podle frekvence).

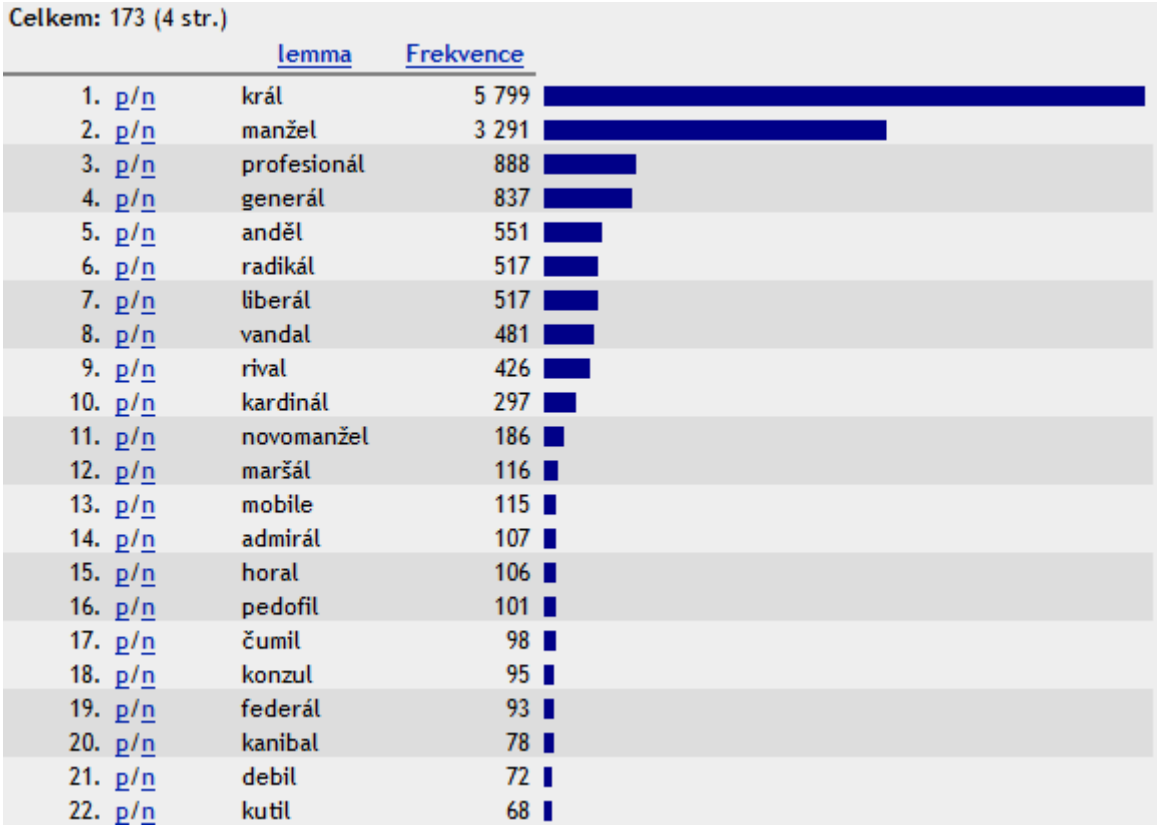
Nicméně propria jako parasystém ponecháme stranou a zaměříme se pouze na apelativa, a to tak že odstraníme lemmata, která začínají velkým písmenem.[6] Seznam lemmat se dále redukuje na 173 (obrázek 8).

Ručním tříděním získáme ze 173 lemmat 40 lemmat, která jsou hledanými substantivy konvertovanými z l-ových příčestí. V korpusu SYN2010 lze uvedeným postupem najít tato lemmata: *kutil, čumil, čmuchal, ožrala, břídil, patolízal, šťoural, žvanil, slídil, střádal, packal, hejkal, reptal, tlachal, rýpal, rejpal, hýřil, chlubil, všeuměl, kýval, kecal, mazal, škrabal, všudybyl, skuhral, loudil, fňukal, brblal, šukal, škudlil, šeptal, čmáral, zlobil, koktal, mlsal, klábosil, hrabal, hloubal, cmrndal, cintal* (seřazeno podle frekvence).

Můžeme tedy konstatovat, že ačkoliv více než ¾ (77 %) dokladů jsou případy, které spadají pod pojem přegenerovávání (174-40=134), tak pokud bychom byli zůstali u prvního pokusu, pak by byla míra přegenerovávání vyšší než 98 % (40 lemmat bychom museli ručně vybrat ze seznamu 3023 lemmat). Míru přegenerování jsme tudíž snížili o 21 %.

[6] Propria lze odstranit pomocí volby negativního filtru vybrat a odstranit lemmata začínající na velké písmeno, tedy [lemma="[AÁBCČDĎFGHIÍJKLMNŇOÓPQRŘSŠTŤUÚVWXYÝZŽ].*"].

Obrázek 8



Celkem: 173 (4 str.)

		lemma	Frekvence
1.	p/n	král	5 799
2.	p/n	manžel	3 291
3.	p/n	profesionál	888
4.	p/n	generál	837
5.	p/n	anděl	551
6.	p/n	radikál	517
7.	p/n	liberál	517
8.	p/n	vandal	481
9.	p/n	rival	426
10.	p/n	kardinál	297
11.	p/n	novomanžel	186
12.	p/n	maršál	116
13.	p/n	mobile	115
14.	p/n	admirál	107
15.	p/n	horal	106
16.	p/n	pedofil	101
17.	p/n	čumil	98
18.	p/n	konzul	95
19.	p/n	federál	93
20.	p/n	kanibal	78
21.	p/n	debil	72
22.	p/n	kutil	68

Problém podgenerování

Otázkou zůstane, jak velké množství lemmat tímto postupem zachyceno nebylo. Pomineme případy, kdy pádová substantivní koncovka je homonymní s rodovou koncovkou l-ového příčestí (tvarová homonymie *kutil, kutila, kutily),* které většinou vykazují velké množství chyb v disambiguaci, takže jsme je ze své analýzy záměrně vyloučili, a zaměříme se na případy nezaznamenané ve slovníku morfologického analyzátoru, tedy „odpadkový koš“ (tag=„X.*“). Zopakujeme dotaz, ale místo maskulin životných budeme hledat slova s označením slovního druhu „X“.

V cql bude dotaz vypadat takto: **[word=".*[aáeěiuy]l((ové)|(ů)|(ům)|(ech)|(e)|(ovi))" & word!=".*[bpfvmdtnkghr]el((ové)|(ů)|(ům)|(ech)|(e)|(ovi))" & word!=".*[aáeěéiíoóuůyý][aáeěiuy]l((ové)|(ů)|(ům)|(ech)|(e)|(ovi))" & tag="X.*"]**

Po odstranění proprií[7] získáme ručním tříděním dalších 10 lemmat z celkového počtu 731 výskytů (*plazil, velebil, remcal, plížil, patlal, muchlal, kousal, drnkal, cachtal, brouzdal*).

[7] Viz předcházející poznámka.

Nejde jen o hapaxy (obrázek 9), ačkoliv je třeba připomenout, že doklady pocházejí z jediného dokumentu.

Obrázek 9

Výskytů: 18 i.p.m.: 0,15 (vztaženo k celému korpusu) | ARF: 1

Hvězdotřesení	. Okamžitě ho obklopili tři	plížilové	/plížilové/X@----------	, půl tuctu plížovníků a
Hvězdotřesení	se neboj . Ti hluční	plížilové	/plížilové/X@----------	už jsou pryč . "
Hvězdotřesení	všechen ten rámus . Ti	plížilové	/plížilové/X@----------	dokážou rozvibrovat celý blok .
Hvězdotřesení	patrně důvodem , proč byli	plížilové	/plížilové/X@----------	nejoblíbenějšími mazlíčky čílů . Skoro
Hvězdotřesení	Tekutý písek . " Ostatní	plížilové	/plížilové/X@----------	budou rádi , že zase
Hvězdotřesení	. Pokud vejce nesežrali toulaví	plížilové	/plížilové/X@----------	, mělo mládě slušnou šanci
Hvězdotřesení	okusoval trávník , Valík a	plížilové	/plížilové/X@----------	nebyli nikde v dohledu .
Hvězdotřesení	" Tihle geneticky čistí laboratorní	plížilové	/plížilové/X@----------	vypadají všichni stejně , "
Hvězdotřesení	kůži pokrytou puchýři . Domácí	plížilové	/plížilové/X@----------	na tom byli stejně jako
Hvězdotřesení	jí podobné podnikali lovy divocí	plížilové	/plížilové/X@----------	. Někdo o ni pečoval
Hvězdotřesení	očima stačil všimnout , že	plížilové	/plížilové/X@----------	se vyděšeně rozprchli už dávno
Hvězdotřesení	v horách , kde se	plížilové	/plížilové/X@----------	chovat nesmějí . Bojím se
Hvězdotřesení	Čílové z okolních klanů i	plížilové	/plížilové/X@----------	zažili už tucty podobných přistání
Hvězdotřesení	" " Tady jsou imperátorovi	plížilové	/plížilové/X@----------	chránění před divokými mršty .
Hvězdotřesení	ani jeden nechybí . Imperátorovi	plížilové	/plížilové/X@----------	tu mají víc krmení .
Hvězdotřesení	Imperátorovi náleží vše . Všichni	plížilové	/plížilové/X@----------	, každičký korální ořech ,
Hvězdotřesení	, kde jsou chlupatí jako	plížilové	/plížilové/X@----------	. " " To je
Hvězdotřesení	jako by to byli nějací	plížilové	/plížilové/X@----------	v zoologické zahradě ! "

Optimalizaci míry podgenorávání spatřujeme v automatizovaném vytvoření seznamu kandidátů na hledané jednotky z tvarů nezachycených slovníkem automatického morfologického analyzátoru (tvarů, které splňují podmínky otestované na datech rozpoznaných automatickou morfologickou analýzou a zároveň mají značku tag=„X.*“). Ručním tříděním automaticky pořízených dat se podařilo získat celkem dalších 10 lemmat, která nebyla zachycena výše uvedeným postupem a spadala tak pod pojem podgenerování. Celkový seznam 40 lemmat se takto rozšířil o dalších 10 lemmat (zlepšení o 25 %). Efektivnost uvedeného vyhledávání lze porovnat také s postupy popsanými jinde (srov. Štícha 2011 : 225–226).

Závěr

Ukázali jsme, jak lze při zadání dotazu korpusovému manažeru na základě pozorování přegenerovaných dat postupně optimalizovat dotaz tak, aby se počet dokladů, jež je dále třeba podrobit ruční analýze, pokud možno minimalizoval. V uvedeném příkladu se nám podařilo snížit míru přegenerovávání o 21 %. Vyzkoušený postup jsme zopakovali na datech, která nejsou rozpoznána automatickou morfologickou analýzou a spadají tudíž pod pojem

podgenerovávání. Ruční analýzou takto získaných dat se nám podařilo rozšířit počet lemmat o 25 %.

Redukce počtu lemmat pro ruční analýzu přispěla k větší efektivitě a spolehlivosti vyhledání maximálního počtu jednotek; minimalizace problému přegenerovávání i podgenerovávání vede obecně k redukci rozsahu ruční práce, která je vždy nákladná (na čas popř. i finanční prostředky, z nichž je třeba hradit školené anotátory) a představuje také zvýšené nebezpečí chyb z nepozornosti, které jsou následně poměrně obtížně detekovatelné.

Uvedený postup je případem metody korpusové lingvistiky, kdy užíváme korpus nejen jako zdroj pro lingvistické observace (corpus based), ale též jako zdroj odhalování pravidel fungování přirozeného jazyka (corpus driven), která mohou být úspěšně využita i k jiným účelům, než jsou ty, jimž slouží v daném případě. Konkrétně si lze představit např. využití uvedených pravidel při tvorbě guesserů (hadačů), tedy programů, které se na základě různých technik snaží „uhádnout" vlastnosti jednotek přirozeného jazyka (např. interpretace lemmatu a morfologického tagu).

Bibliografie

DOKULIL, M. a kol. (1986): Mluvnice češtiny 1. Praha: Academia.

KOMÁREK, M. a kol. (1987): Mluvnice češtiny 2. Praha: Academia.

McENERY, T. – HARDIE, A. (2012): Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press.

OSOLSOBĚ, K. (2011): Morfologie českého slovesa a tvoření deverbativ jako problém strojové analýzy češtiny. Brno: Masarykova univerzita.

OSOLSOBĚ, K. (2011): Korpus jako zdroj dat pro studium slovotvorby. In V. Petkevič, – A. Rosen (eds.), 3. Gramatika a značkování korpusů. Praha: Nakladatelství Lidové noviny/Ústav Českého národního korpusu, s. 10–23.

ŠTÍCHA, F. (ed) (2011): Kapitoly z české gramatiky. Praha: Academia.

Elektronické zdroje

Korpus SYN2010: Český národní korpus - SYN2010. Ústav Českého národního korpusu FF UK, Praha 2010. Cit. 6. 11. 2013, dostupný z WWW: <http://www.korpus.cz>.

Korpus SYN: Český národní korpus - SYN. Ústav Českého národního korpusu FF UK, Praha 2010. Cit. 6. 11. 2013, dostupný z WWW: <http://www.korpus.cz>.

Korpus czTenTen12. Cit. 6. 11. 2013, dostupný z WWW: <https://ske.fi.muni.cz>.