SBORNÍK PRACÍ FILOZOFICKÉ FAKULTY BRNĚNSKÉ UNIVERZITY
STUDIA MINORA FACULTATIS PHILOSOPHICAE UNIVERSITATIS BRUNENSIS
A 56, 2008 — LINGUISTICA BRUNENSIA

KLÁRA OSOLSOBĚ

FORMÁLNÍ PRAVIDLA DERIVACE DEVERBATIV NA -*Č*

Úvod

Cílem studie je otestovat možnosti algoritmizace slovotvorných postupů a ověřit tak možnosti automatizace vytváření alespoň některých částí derivačního slovníku češtiny. Prostředkem ke zvolenému cíli bude rozbor jednoho derivačního typu jmen odvozených od sloves sufixem *-č.* Při analýze budeme používat webové rozhraní *Deriv* vytvořené na FI MU*,* které pro generování možných derivačních vztahů používá slovník automatického morfologického analyzátoru *ajka* (Sedláček, 2004), a data získaná z korpusů ČNK (SYN2000, SYN2005, SYN2006PUB). Pokusíme formulovat pravidla použitelná jak při extrakci dat ze slovníků morfologického analyzátoru *ajka* prostřednictvím *Deriv,* tak při formulaci dotazů pro získání příslušného materiálu z jazykových korpusů.

Výsledkem by měla být a) formulace formálních pravidel pro derivaci deverbativ tvořených sufixem *-č*, b) vytvoření co nejrozsáhlejšího seznamu torzovitě popsaných popřípadě málo doložených a ve slovnících nezachycených výjimek a anomálií, který pomůže zpřesnit empiricky navržená pravidla, c) vytvoření pokud možno co nejrozsáhlejšího seznamu možných homonymií generovaných automaticky při aplikaci navržených pravidel.

Ajka Derivační rozhraní

Deriv (http://nlp.fi.muni.cz/projekty/ajka/cjbb85/) je nástroj vyvinutý na FI MU, který umožňuje jednoduchým způsobem vyhledávat ve strojovém slovníku automatického morfologického analyzátoru *ajka* (Sedláček, 2004) lemmata podle formálně zadaných pravidel (úvodní řetězec, koncový řetězec, morfologická značka) a vytvářet touto cestou seznamy slov, která s velkou mírou pravděpodobnosti patří k jednomu derivačnímu typu, respektive jsou tvořena jedním derivačním prostředkem. Se seznamy lze následně pracovat, je možné je prohlížet ve dvou modech, a to jako prosté seznamy, nebo jako seznamy s uvedením frekvence vyhledaných jednotek v korpusu SYN2000 (viz níže). Nástroj tedy umožňuje

rychlé prohledání a extrakci dat z rozsáhlého strojového slovníku českých kmenů (Osolsobě, 1996, Sedláček, 2004), je propojen s korpusem SYN2000 (100 milionů slovních tvarů), a umožňuje tak v korpusu ověřovat frekvenci jednotek zahrnutých ve strojovém slovníku.

Poznámka: Strojový slovník češtiny (Osolsobě, 1996) byl vybudován na základě hesláře SSJČ a doplněn o řadu dalších slov na základě testování aplikací automatických morfologických analyzátorů *lemma* (Ševeček, 1996) a *ajka* (Sedláček, 2004) na korpusech češtiny (korpusy ČNK, korpusy budované na FI MU). Strojový slovník analyzátoru *ajka* zahrnuje zhruba 400 000 jednotek (kmenů), k nimž na základě formálních pravidel (deklinačních vzorů) generuje 6 milionů slovních tvarů.

2. Analýza materiálu

V tomto oddíle ukážeme, jak lze využít nástroje *Deriv* ke zjišťování relevantních informací o dvojicích sloveso – potencionální odvozené substantivum.

Po přihlášení (uživatelské jméno/heslo) v *Deriv* zvolíme z nabídky funkci **Hledání slov podle pravidla**. Nejdříve zadáme do nabídky příkaz pro vyhledání slov končících na *-t* a majících značku k5.* (sloveso) a uložíme je do souboru (1). Pak zadáme příkaz pro vyhledání slov končících na *-č* a majících značku k1gM.* (substantivum, maskulinum životné) a uložíme je do souboru (2).

Vytvořené soubory (1) a (2) sloučíme a vytvoříme z nich soubor jeden (3). Na takto vzniklý soubor použijeme funkci **Hledej základová slova** a zadáme, že chceme v příslušném souboru najít dvojice slov takových, že jeden člen páru končí na *-t* a druhý je slovo vzniklé odtržením *-t* a jeho nahrazením *-č*. Dále aplikujeme funkci **Rozdělení souboru** a získáme dva soubory, jeden, který obsahuje nalezené dvojice (4) a druhý, který obsahuje zbylá slova, ke kterým se nepodařilo najít podle příslušného pravidla slovo do „páru“ (5).

Soubor (4) pak obsahuje kandidáty na hledaný derivační typ.

Poznámka: Přegenerování (případům, které jsou z hlediska automatické morfologické analýzy homonymní s tvary generovanými podle zadaného pravidla) dvojic typu **klít/klič, *mít/míč* a **mlít/mlič* se vyhneme zadáním morfologické značky (k1gM.* - maskulinum životné), dvojici *kout/kouč* je třeba zařadit k výjimkám.

Popsaným postupem získáme seznam, který zahrnuje 720 dvojic. Z dvojic sloveso – činitelské jméno nezahrnuje ty, pro které neplatí jednoduše formulované vyhledávací pravidlo, tedy z lingvistického hlediska případy, kdy formantem činitelského jména není pouze sufix *-č* a odvozovacím základem není kmen (! infinitivní) /kořen slovesa, ale kdy při derivaci dochází k hláskovým alternacím, a to buď k alternacím kořenového vokálu (KoS), anebo kmenového vokálu (KmS).

Poznámka: Přegenerovávání v případě dvojice *halit/halič* je třeba přičíst na vrub chybě morfologického analyzátoru *ajka* (substantivum *halič* (!*h*) se interpretuje jako maskulinum životné. U dvojice *kočkot* (*k1gM.**) je chybou derivační-

ho nástroje, který při aplikaci funkce **Rozděl soubor** nepracuje s morfologickými značkami. Totéž platí o dvojici *myč/myt,* (*myt* - tvar pasivního t-ového participia), jimž bychom se mohli vyhnout při aplikaci funkce **Hledání slov podle pravidla** specifikací značky *k5.***mF* (infinitiv).

2.1 Alternace KoS a KmS a možnosti jejich vyhledávání - *Deriv*

V této části přihlížíme ke studii (Klímová, Štícha, 2006). K popisu derivace deverbativ na *-č* uvedenému v této studii jsme na mnoha místech navrhli úpravy a doplnění.

2.1.1 Samohláskové alternace KmS

Pro vyhledávání dvojic s alternací KmS lze pro automatický analyzátor formulovat příslušná pravidla poměrně jednoduše, protože alternující vokál předchází bezprostředně před sufixem *-č*, a při vyhledávání podle koncového řetězce znaků lze zadat místo jednoho znaku (*-t* > *-č*) znaky dva. Sledujeme-li systém českých slovesných tříd, pak zjistíme, že lze doložit pravidelné alternace kmenového vokálu *-a-t/-á-č* u několika sloves první třídy podle kmene přítomného (vzory *brát* a *mazat : orat/oráčt, mazat/mazáč*), a u sloves páté třídy podle kmene přítomného vzoru *dělat/kopat* (*kopat/kopáč*). Samohláskovou alternaci KmS má substantivum derivované od slovesa *mlít*, které bývá zařazováno buď jako nepravidelné (Šmilauer, 1971), nebo jako sloveso první třídy podle kmene přítomného vzor *umřít* (Komárek, 1986). Alternace jsou též doloženy u malé skupiny nepravidelných sloves čtvrté třídy podle kmene přítomného s dlouhou KmS *-í-t/-i-č* (*mstít/mstič*).

2.1.2 Samohláskové alternace KoS

Formulace pravidel pro automatický analyzátor je poměrně snadná v případě, že kořenový vokál bezprostředně předchází před sufixem *-č*, tedy z lingvistického hlediska v případě sloves třetí třídy podle kmene přítomného vzory *krýt* (*mýt/myč*).

Poměrně složitější je formulovat pravidla alternací KoS v případě, že KoS bezprostředně nepředchází před sufixem *-č*.

Jde o některá slovesa 4. třídy vzor *prosit* : typ *loupit/lupič,* o slovesa vzoru *sázet*, která mívají dublety : typ *sázet/s(a|á)zeč* a o slovesa 1. a 5. třídy vzorů *mazat/brát* a *dělat*, která mívají rovněž dublety : typ (a) *vázat/vazač* / *vzpírat/vzpěrač* a typ (b) *uspávat/usp(á|a)vač.*

2.2 Typy alternací KmS a KoS stojících bezprostředně před sufixem *-č*

2.2.1 Alternace KmS .*(a|e|ě|i|í)t.>.*(á|e|ě|i)č

Poznámka: Nadále budeme užívat regulární výrazy obvyklé pro práci s korpusovými manažery (např. Bonito, Rychlý, 2000) i jinými aplikacemi. *Deriv*, funkce **Hledání slov podle pravidla** pracuje se zadáním, které používá regulární výrazy. Znak „.“ se užívá pro libovolný znak (písmeno), znak „*“ za libovolné opakování předchozího znaku, znak „|“ znamená „nebo“ (disjunkce), znak „&“ znamená „a současně“ (konjunce).

.*at/.*áč	.*ít/.*eč	.* (e|ě)t/.*ič	.* it/.*(e|ě)č	.*ít/.*ič
kopat/kopáč	*mlít/mleč*	*držet/držič*	*obalit/obaleč*	*mstít/mstič*
*22-4=18**	*1***	*1****	*1*****	*1*

* Poznámka: Čtyři automaticky generované případy nelze pokládat za deriváty sloves, nýbrž jsou odvozeny od substantiv : **klepetat/klepeto/klepetáč, *krkat/krk/krkáč* (z něm. Geizhals), **volat/vole/voláč*, a patří tedy do kategorie substantiv podle význačné části, nebo od adjektiva (**zelenat/zelený/zelenáč*), patří do kategorie nositel vlastnosti.

** Poznámka: Chybám typu **klít/kleč, *lít/leč*, **sít/seč* se vyhneme zadáme-li morfologickou značku (k1gM.* - maskulinum životné). Zabráníme tak mylnému vyhledání feminin, která jsou homonymní s automaticky generovatelnými deverbativy. Analyzátor nevyhledá dvojici *plít/pleč* (srv. níže).

*** Poznámka: Zdá se, že od sloves 4. třídy podle kmene přítomného vzoru *trpět* se deverbativa na *-č* téměř netvoří.

**** Poznámka: Tento typ doložený pouze jediným dokladem je sporný proto, že za fundující sloveso substantiva *obaleč* se pokládá *obalit*, tedy sloveso dokonavé (srv. níže).

2.2.2 Alternace KoS .*(í|ý)t.>.*(i|y|e)č

.*ít/.*ič	.*ýt/.*yč	.*ít/.*eč
šít/šič	*mýt/myč*	
*3-2**	*1*	*0*

* Poznámka: Zařazení substantiva *dojič* a jeho domnělá fundace verbem *dojít* nás vedla k následující úvaze. Domněle fundující sloveso *dojít* je slovesem dokonavým. Deverbativa kategorie činitelských jmen tvořená sufixem -č jsou v naprosté většině případů tvořena od sloves nedokonavých (srv. Šmilauer, 1971, s. 27, Dokulil, 1986, s. 237). Zdá se, že by bylo možné zabránit automatickému generování chybné interpretace zavedením pravidla, které by vyloučilo práci se slovesy dokonavými (konkrétně při výběru slov na *-t* ze slovníku *ajky* vybrat podle značky pouze slova se značkou *k5aImF* – infinitivy nedokonavých sloves). Rozbor feminin (?přechýlených jmen činitelských např. *hráčka, šička*, *plečka* a ženských názvů prostředků např. *hračka, pračka, plečka*) jsme v této studii ponechali stranou, ačkoliv jde o velmi zajímavou problematiku zejména z hlediska hláskových alternací.

2.3 Typy alternací KoS nestojících bezprostředně před sufixem *-č*

2.3.1 Typy alternací KoS sloves na *-it*

.*á.it/.*a.ič	.*í.it/.*i.ič	.*ou.it/.*u.ič	.*ý.it/.*y.ič	.*í.it/.*ě.ič
pálit/palič	*řídit/řidič*	*loupit/lupič*		
*5**	*2*	*6*	*0***	*0***

* Poznámka: Započítány jsou i případy **.*á..it/.*a..ič** *(dlážďit/dlaždič)*.

** Poznámka: Vzhledem k tomu, že existují slovesa *bílit* i *bělit* a činitelská jména *bílič* i *běličí*, dáváme pro synchronní popis jazyka přednost interpretaci *bílit/bílič* a *bělit/běličí* a nepředpokládáme alternace KoS **.*í.it/.*ě.ič**. Pro automatickou morfologickou analýzu (a nejen pro ni) jsou problematické dvojice sloves inchoativum/faktitivum např. *tupět/tupit,* jejichž významový rozdíl se často stírá (srv. Šmilauer, 1971, s. 157, a také interpretaci slovesa *bělet* v SSJČ : „*bělet* 1. …*stávat se bílým* …; 2. *(co) činit bílým*: kávu b. mlékem…“). Činitelská jména se logicky tvoří od faktitiv. Rovněž nebyly doloženy alternace typu **.*ý.it/.*y.č** (např. sloveso *hýřit/*hyřič*).

2.3.2 Typy alternací KoS sloves na souhlásku *-t*

.*áCt/.*a.eč	**.*éCt/.*e.ač**
klást/kladeč	*plést/pletač*
1	*1*

Poznámka: „C“ užíváme za libovolný konsonant, tedy (b|c|č|d|f|g|h|j|k|l|m|n|ň|p|q|r|ř|s|š|t|ť|v|w|x|z|ž), skutečně doloženy jsou pouze případy (c|s|z). Jde o deverbativa od sloves 1. třídy podle kmene přítomného vzorů *nést/péct.* Jak dokládají další (i neživ.) deverbativa na *-č*, samohláskou vkládanou mezi kmen zakončený na souhlásku a sufix *-č* může být jak *-a-* nebo *-e-* (viz výše), tak *-á-* (srv. např. *péct/pekáč* - .*éCt/.*e.áč, nebo *?sícť(sekat)/sekáč* - .*íCt/.*e.áč). Dokladem pestrosti alternujících KoS může být i *hnísť/hnětač* (.*íCt/.*ě.ač).

2.3.3 Typy alternací KoS sloves na *-(e|ě)t*

.*á.(e|ě)t/.* (a|á).(e|ě)č	**.*í.(e|ě)t/.* (i|í).(e|ě)č**
sázet/s(a|á)zeč	***odklízet/odkl(i|í)zeč***
16+1*	***1***

* Poznámka: Fundace substantiva *povaleč* perfektivním slovesem *poválet* (perfekt.) je sporná (srv. výše).

Poznámka: U deverbativ s KoS *-ou- (vysoušeč, ostouzeč, zkoušeč, pokoušeč, spouštěč, ...)* nebyl doložen případ alternace. Rovněž jsme nenašli případ alternace KoS *.*í. (e|ě)t/.*(e|ě).(e|ě)č.*

2.3.4 Typy alternací KoS sloves na *-at*

.*á.at/.*a.ač	**.*í.at/.*ě.ač**	**.*í.at/.*i.ač**	**.*ou.at/.*u.ač**	**.*ý.at/.*y.ač**
vázat/vazač	***vzpírat/vzpěrač***	***stříhat/střihač***	***foukat/fukač***	***ohýbat/ohybač***
27*	***24***	***9***	***7****	***4***

* Poznámka: V uvedeném počtu jsou zahrnuta pouze slovesa, která nekončí na *-ávat*.

** Poznámka: Započítány jsou i případy **.*ou..at/.*u..ač** *(poslouchat/posluchač)*.

3. Porovnání možností a výsledků *Deriv* a korpusových sond

Deriv je poměrně výkonný nástroj, odhalí však pouze případy, kdy jsou oba členy vyhledávané dvojice uvedeny ve strojovém slovníku kmenů. Sondy do korpusů ovšem ukazují, že by bylo žádoucí, aby bylo možné pracovat s potencionálně generovatelnými tvary (pravidly pro odvozování příslušných dvojic) a s korpusy jako zdroji dat pro ověřování existujících derivátů. V následující kapitole proto ukážeme bohatství existujících korpusů. Ukážeme, jak je možné poloautomaticky (použitím formálních pravidel) vyhledávat v korpusech kandidáty na analyzovaný derivační typ, a to především mezi slovy, kterým automatická morfologická analýza nepřiřadí správnou značku.

4. Korpusy ČNK

Jazykové korpusy představují v současnosti uznávaný a preferovaný zdroj dat pro nejrůznější lingvistické observace. V této studii budeme sledovat, jak mohou korpusy psaného jazyka (češtiny) pomoci při přípravě podkladů pro pravidla automatické derivace zvoleného typu. Budeme pracovat s korpusy SYN2000, SYN2005 a SYN2006PUB. Pokusíme se ukázat, jakým způsobem lze z lemmatizovaných a morfologicky anotovaných korpusů shromáždit co nejrozsáhlejší materiál pro studium vybraného derivačního typu (odd 4.1) a jak lze zužitkovat i nedostatky automatické morfologické analýzy (odd 4.2).

4. 1 Pravidla pro vyhledávání

Sledované korpusy jsou anotovány na rovině tzv. morfologického značkování a jsou lemmatizovány (Hajič, 2004). Uvedeme systematický postup vyhledávání kandidátů derivačního typu jmen činitelských tvořených sufixem *-č* v korpusech označkovaných automatickým analyzátorem J. Hajiče (dále *HA*).

Pokud tedy chceme tyto kandidáty v korpusu vyhledat, pak můžeme zvolit následující postup:

Vyjdeme z předpokladu, že činitelská jména tvořená sufixem *-č* jsou **maskulina životná**. Tím je odlišíme od slovotvorně stejně tvořených a někdy homonymních jmen prostředků (např. *nosič*). Do dotazovacího řádku zadáme následující dotaz.

[lemma=".*č" & tag="NNM.*"]

Prozkoumáme-li výsledky (statistický přehled lemmat), pak můžeme na základě empirických dat formulovat **pravidlo č. 1:** ***Pokud u substantiva maskulina životného končícího na -č předchází před -č souhláska, pak se nejedná o činitelské jméno.***

Použitím negativního filtru odstraníme konkordanční řádky.

N/filtr

[lemma=".*(b|c|č|d|d'|f|g|h|j|k|l|m|n|ň|p|q|r|ř|s|š|t|t'|v|w|x|z|ž)č"]

(Odstraníme případy jako např. *Renč, pinč, komanč, ...*)

Dále můžeme formulovat **pravidlo č. 2:** ***Pokud u substantiva maskulina životného končícího na -č předchází před -č samoláska o-, ó-, u-, ů-, (ú-), é- , pak se nejedná o činitelské jméno.***

Použitím negativního filtru odstraníme konkordanční řádky.

N/filtr

[lemma=".*(o|u|ů|ú|ó|é)č"]

(Odstraníme substantiva jako *Bezruč, kouč, Koč, roztoč, ...*)

Empiricky lze dále zjistit, že soubor obsahuje množství proprií (zejména jména na *-ič*), která nejsou deverbativy. Apelativa (tedy i jména činitelská na *-č*) mohou fungovat jako propria (srv. např. příjmení typu *Mazáč*, ...).

Ruční analýzou seznamu, který jsme získali použitím p-filtru pro vyhledání slov, jejichž lemma je velké písmeno (propria), jsme získali 590 lemmat. Pouze k několika (např. *Tkáč, Klapáč, Klepáč, Kováč, Mazáč, Řezáč, ?Radič, ?Prudič, Hajič, ?Bulič, ?Pulič, ?Zvonič*) by bylo možné tvořit česká slovesa. Problematiku vlastních jmen ponechává tato studie stranou. Přistoupili jsme tudíž k odstranění proprií, přičemž jsme vyšli z lemmatizace (lemma začíná velkým písmenem).

N/filtr

[lemma="(A|Á|B|C|Č|D|Ď|E|É|F|G|H|I|Í|J|K|L|M|N|Ň|O|Ó|P|Q|R|Ř|S|Š|T|Ť|U|Ú|V|W|X|Y|Z|Ž).*č"]

(Odstraníme vlastní jména, především jihoslovanská na -ič : *Karadžič, ...*)

Na základě další analýzy empirických dat můžeme formulovat **pravidlo č. 3.:** ***Pokud u substantiva maskulina životného končícího na -č předchází před -č samoláska á-, jde poměrně často o desubstantiva (jména podle význačné části/znaku) typ hlaváč, nebo o deadjektivní názvy nositelů vlastností typ zelenáč.***

Pomocí pozitivního filtru lze vytvořit seznamy, ty je pak třeba ručně analyzovat.
P/filtr
[lemma=“.*áč“]
Konkordance>Statistiky>Frekvenční distribuce

	.*áč	činitelská	ostatní
SYN2000	56	17*	39
SYN2005	50	29**	21
SYN2006PUB	58	20***	38

**hráč/32016, spoluhráč/2210, protihráč/447, rváč/209, spáč/198, kopáč/164, sekáč/108, oráč/70, voráč/67, štváč/55, sráč/51, česáč/14, kokrháč/6, srdcerváč/4, řezáč/4, práč/3, zváč/2*.
Poznámka: V případě deverbativa sekáč upozorňujeme na možnou homonymii v rodě (*sekáč* - prostředek např.: *na tu že by musel <**sekáčem**> a kladivem*, i na možnou homonymii *seconhand-sekáč* např.: . ... *sekáči kupují v <**sekáči**>...*) První z nich automatická morfologická analýza reflektuje, výsledky disambiguace příslušné homonymie nejsou ovšem příliš uspokojivé. Druhá na rovině morfologické analýzy pochopitelně reflektována není. Substantivum *voláč* je termínem z oblasti zoologie *holub voláč* a jde o desubstantivum (od *vole, -ete*) nikoli o deverbativum (srv. např. SSJČ, PSČ). Substantivum *zelenáč* není derivováno od slovesa *zelenat se*, ale od adjektiva *zelený.*
***hráč/16822, spoluhráč/1673, rváč/331, protihráč/317, spáč/293, kopáč/203, smrkáč/157, oráč/126, sráč/110, sekáč/102, česáč/102, štváč/60, řezáč/41, voráč/28, hrabáč/22, práč/11, srdcerváč/10, žráč/10, makáč/5, kokrháč/5, psáč/2, nahraváč/2, smáč/1, ochutnáváč/1, odzdisráč/1, načesáč/1, exspoluhráč/1, vrzáč/1, jakobyspoluhráč/1.*
Poznámka: V případě substantiv *načesáč/1* a *jakobyspoluhráč/1* jde o okazionalismy, jak vysvíta z kontextu *(...Na počátku byl sen a slova „naháči a <**načesáči**>“, která napadla spisovatele Františka Zborníka..., ...že své < **jakobyspoluhráče** > nesekal hokejkou...)*. Substantivum *poslucháč(5)* pochází ze slovensky psaných textů. U substantiv *nahraváč (2), ochutnáváč (1)* jde patrně o překlepy nikoli o alternace KmS. U substantiva *visáč(1)* z kontextu plyne, že jde o univerbizaci za *visací zámek* mylně analyzovaný automatickou morfologickou analýzou jako tvar maskulina životného. Zdá se, že automatická morfologická analýza opřená o přesnější popis pravidel derivace by neměla generovat k substantivu *visáč* fundující sloveso (*viset/*vis(í|e)č* nebo **visat/visáč*). Substantiva *klepetáč* a *voláč* (viz výše) nejsou deverbativy (nejsou odvozena od verb *klepetat, volat*). O substantivech *sekáč, zelenáč* platí to, co bylo řečeno v předešlé poznámce.
****hráč/160675, spoluhráč/3880, protihráč/2320, rváč/1047, kopáč/563, spáč/458, oráč/258, sekáč/201, sráč/173, voráč/151, štváč/113, řezáč/74, česáč/51, superhráč/27, práč/22, srdcerváč/19, hrabáč/14, smrkáč/9, kokrháč/8, zváč/1*.
Poznámka: O substantivech *sekáč, zelenáč* a *voláč* platí to, co bylo řečeno v předešlých poznámkách.
V následující tabulce uvádíme počty kandidátů na činitelská jména (KČJ) v jednotlivých sledovaných korpusech, počty lemmat na *-č*, substantiv, která nelze interpretovat jako činitelská jména (NČJ) získané ruční analýzou, počty činitelských jmen na *-áč* získaných ruční analýzou (viz výše) a celkový počet správně lemmatizovaných (maskulina životná) jmen činitelských na *-č* (kromě proprií).

	KČJ	NČJ	činitelská na -áč	CELKEM ČJ
SYN2000	374	4*	17	387
SYN2005	738	73**	29	694
SYN2006PUB	449	6***	20	462

** carevič, čuvač, knírač* a proprium *Dědič.*
Poznámka: Váhali jsme se zařazením substantiva *trubač*. V MČ 1 (srv. Dokulil, 1986 s. 237) se uvádí, že : „Přípona *-č* má řídce užívané varianty **-ač**: *trubač, vozač* a –eč: *kladeč*.“ Pokud by mělo být v rámci popisu derivace uvedeno fundující sloveso, museli bychom mít pro uvedené a další případy zvláštní pravidlo pro deverbativa s okrajovou alternací KmS .*it/>.*(a|á)č (*troubit/trubač, vozit/vozač,* obdobně i pro neživ. *kropit/kropáč*).

[**]*carevič, dědič (Dědič), knírač, rač(tvar rači=raději), apač (Apač), čuvač, znič (zničeho/znič nic), vozač, uvnič (místo uvnitř), čeč (člověče/čeč), vůbeč (místo vůbec), pumlič, prič, žeč, novořeč, borisyč (Borisyč), prrryč (pryč), taj-č, brič, 90,90Kč, prostorovič (vlastní jméno - překlep), babič (tvar babičy), vobyč, knihotlač, prostředníč (prostředníče/prostředníč/NNMS1.* správně má být : prostředníče/prostředník/NNMS5.*), hódžič (hódžiča . druh čaje), dutič, zkrenčič, skřenčič, krač, palač, tovnič, před-č, mič (Hoto no mič), pohlčovač, visielač, 900.000Kč, něčeho-něč, mučač, snoubič, mopač, čoveč, mělič, jiného-něč, 14Fryč, bač (tvar i bačov sb. bača), izolač, plyč, slépič, spotykač, šeč, síč, michajlyč, krč (rrr), zarobič, llkovič, basmač, prčič (do prčič), popovič, božič, čača/ča-č, tlačtlač, 150000Kč, tai-č, mitryč, oči-č, pič (vesměs tvary substantiva piča), břitvač, poněvač, neč (tvar nečeho), rapač, tarač, pyč, martinyč (Martinič), martynyč (Martynyč).*

Poznámka: Jak je patrné jde o řadu náhodných překlepů (často ve vlastních jménech), o chyby proti normě naznačující vady výslovnosti, o problematicky tokenizované tvary. Zdá se, že řada mylně zahrnutých tvarů se sem dostala při testování guesserů (viz níže). Otázkou, kterou si klademe, je, kolik z těchto tvarů je takových, že by mohly být deverbativními tvary od běžných českých sloves. Odpověď může být zajímavá proto, že by měla odhalit možné chyby, jichž by se mohl dopustit automatický generátor deverbativ. Mezi výše uvedenými lemmaty jsme empiricky zjistili pouze následující: *dědit/dědič, znít/znič, mít/mič, snoubit/snoubič, bát/bač, sít/síč, prčit/prčič, pít/pič, ?prostřednět/prostředníč.* Ke všem ostatním by nemělo být nalezeno ve slovníku kmenů příslušné sloveso. Případ *zničeho/znič/NNM.** by dle našeho názoru vyloučilo pravidlo, které nepřipouští u příslušného typu zakončení *-eho*. Týž přístup by bylo možné aplikovat u případu *bačov/bač/NNM.**. Ostatní případy lze ovšem pokládat za kandidáty homonym generovaných a potažmo rozpoznávaných automatickou morfologickou analýzou. Chybám jako **prostřednět/prostředníč* by se dalo zabránit aplikací pravidel alternací KmS. Zajímavé je z tohoto hlediska substantivum *snoubič*, o němž z příslušného kontextu vysvítá, že se jedná o slangový výraz derivovaný zkracováním a grafickou adaptací od anglického *snowboard*. Internetové diskuse o tomto slově ovšem svědčí o tom, že i rodilé mluvčí napadne souvislost se slovesem *snoubit*, přestože slovníky (PSJČ, SSJČ) dokládají pouze *snubič* .

[***] *carevič, čuvač, knírač* a propria *Dědič, Gluščevič, Zuvač.*

4.2 Co mohou korpusy doplnit k informacím uváděným ve slovnících a potažmo i v automatických analyzátorech

Korpusy jsou značkovány automatickými morfologickými analyzátory budovanými na základě slovníků kmenů, které se opíraly o dostupné slovníky a další datové zdroje. Řadu slov z periferie slovní zásoby zachycenou ve strojových slovnících analyzátory rozpoznávají. S periferními výrazy nezachycenými ve strojových slovnících pracují různě.

Naším cílem bylo zjistit, jak pracuje automatická analýza se slovy/slovními tvary, které nejsou uvedeny ve slovníku automatického analyzátoru. Zvolili jsme jednoduchý postup. Vyhledali jsme všechna slova, jejichž lemma končí na *.*č* a která nejsou označkována jako maskulina životná a všímali jsme si značek. Kromě jmen prostředků (*tlumič, pohrabáč, zesilovač*), neživotných nositelů vlastností (?*listnáč, pečenáč, uzenáč*) a substantiv nazvaných podle význačné části (*ucháč, ...*), popřípadě desubstantivních názvů prostředků *(květináč, ...)* bylo třeba vyloučit feminina *(řeč, síč, leč, ...),* zájmenné spřežky (*oč, nač, zač, seč, več, ...*), neohebná slova (*pryč, jináč, leč, vniveč, napříč, heč, ...*), zkratky (*vč., tč., kč, Kč, ...*). Kromě řady překlepů (slova spojená, rozdělená a jinak špatně napsaná) obsahoval takto vytvořený seznam předpokládané kandidáty na příslušná deverbativa. V korpusu SYN2000 a v korpusu SYN2006PUB měla tato slova většinou značku *X.**.

Poznámka: Morfologický analyzátor Jana Hajiče (Hajič, 2004) použitý pro značkování ČNK pracuje se značkou XX.*/X@.* - neznámé slovo. Analýza takto označených jednotek může sloužit a slouží k úpravám automatických nástrojů (rozšiřování slovníku kmenů).

Jiná je situace v korpusu SYN2005, kde jsme si v seznamu získaném stejným postupem u slov, která lze pokládat za deverbativa tvořená sufixem *-č,* povšimli poměrně velmi vysoké frekvence značky *Db.** (adverbium nestupňovatelné).

Poznámka: Při značkování korpusu SYN2005 byly použity tzv. guessery, tj. automatické programy zaměřené na „uhadování" informace u slov, kterým automatický analyzátor nepřiřadí ani jednu interpretaci. Tyto automatické nástroje mohou být založeny na různých přístupech (lingvistických, statistických, kombinovaných) (srv. Hlaváčová, 2001).

Seznamy získané popsaným postupem zahrnují v naprosté většině případů jména činitelská a jména prostředků. Častá jsou mezi nimi kompozita. Kromě názvů osob (jména činitelská) a věcí (jména prostředků) se okrajově vyskytly termíny z oblasti zoologie a botaniky. Obvykle jde o organicky tvořená slova, někdy kalky termínů (*controler/řídič, dreamer/snič*), profesionalismy *(kryč)*, slangové výrazy *(pařič, kalič)* a okazionalizmy *(žíč)*.

Druhou největší skupinu tvoří překlepy, mezi nimiž se poměrně často objevují nesprávně tokenizované tvary končící na *Kč* (např. *000Kč*), dále slova rozdělená (např. *bezpeč, nářeč, pokrač, silnič, ...*) a okrajově i slova spojená (např. *zmeškáčeskýhráč*).

Méně se vyskytují slova odvozená od substantivního/adjektivního základu. Mezi nimi jsou dosti časté případy, které lze interpretovat jako univerbizaci, často jde o okazionalizmy a výrazy příznakové (např. *demáč, prdeláč, kafáč, cibuláč ,...*).

Z beletristických textů pocházejí slova, která graficky zachycují vady výslovnosti (např. *žeč, šeč, ...*) a případy infinitivů jiných slovanských jazyků na *-č* (např. *rozstřelač, oddač, ...*).

4.3 Korpusy a slovníky (slovníky automatických morfologických analyzátorů nevyjímaje)

V rámci této studie není možné popsat vše, co by bylo možné z korpusů získat například pro lexikografické účely. Sonda přináší celou řadu ústrojně utvořených pojmenování, která nejsou zařazena ve slovnících automatických morfologických analyzátorů a řada z nich nefiguruje ani ve slovnících tištěných. V korpusu je lze najít v kontextu, z něhož je v naprosté většině případů dobře patrný jejich význam a z něho plynoucí morfologické vlastnosti (rozdílu jméno činitelské/ název prostředku odpovídá deklinace podle životného/neživotného typu).

V tomto článku se omezíme na to, jak mohou korpusové doklady přispět pro upřesnění formulace pravidel pro automatickou derivaci deverbativ sledovaného typu. Zdá se, že jednou z možností jak rozšířit pokrytí automatické morfologické analýzy odvozených slov by mohla být aplikace slovotvorných pravidel, která by přispěla k identifikaci a rozpoznávání slov, která automatický morfologický analyzátor neidentifikuje na základě dat uložených ve slovníku. Sonda do tří korpusů ukazuje, že ačkoliv řada sledovaných deverbativ jsou slova s malou frekvencí, jedná se slova pravidelně derivovaná od obvyklých a automatickou analýzou rozpoznávaných (slovesných) základů. Pro automatickou morfologickou analýzu by

mohla být podnětná i řada kompozit, u nichž pak zajímavý materiál skýtají především složeniny, kde je hranice složek komponovaného tvaru naznačena graficky (spojovník, lomítko), nebo případy, kdy jde o grafickou chybu a z ní plynoucí nesprávnou tokenizaci.

V tomto článku bychom rádi upozornili především na doklady deverbativ od sloves, u nichž může docházet k alternacím KoS a KmS. Dokladů na tento typ alternací není ani v tištěných slovnících, ani ve slovnících automatických analyzátorů mnoho (viz výše). Korpusy se tak mohou stát cenným zdrojem dat pro přesnější formulaci pravidel na straně jedné a pro případné doplnění seznamů výjimek a anomálií na straně druhé.

Porovnáme-li údaje o jednotlivých typech alternací KmS a KoS získané použitím nástroje *Deriv* (viz výše oddíl 2) s korpusovými daty, můžeme doplnit další příklady jednotlivých alternačních typů.

.*at/.*áč*	.* (e|ě)t/.*ič**	.*ít/.*ič***

*K typu .*at/.*áč *kopat/kopáč* můžeme na základě analýzy korpusových dat doplnit deverbativa *načesáč, makáč, ometáč, vometáč, (* neživ. *hrkáč, brnkáč*). Kromě toho lze na základě korpusů do seznamu výjimek zapsat případy potencionálních homonymií generovaných při aplikaci formálních pravidel. Patřily by sem výše uvedené případy (*voláč/vole/*volat, sekáč/seconhand/*sekat*) a další doložené v korpusech (*rubáč/rub/*rubat*, *tutáč/tuty/*tutat*).

**K okrajově doloženému typu .* (e|ě)t/.*ič *držet/držič* jsme v korpusu SYN2005 našli doklad *prdět/prdič.*

***K typu .*ít/.*ič *mstít/mstič* jsme v korpusech nalezli deverbativum v doloženém případě neživotné (*...na poličce jsem nedávno zahlédl i elektronický <snič> ...) snít/snič* (kalk z angl. *dreamer*).

.*á.it/.*a.ič*	.*í.it/.*i.ič**	.*ou.it/.*u.ič	.*ý.it/.*y.ič	.*í.it/.*ě.ič

*K typu .*á.it/.*a.ič byly v korpusech nalezeny doklady *vážit/važič, hlásit/hlasič.*

**K typu .*í.it/.*i.ič doklad *mířit/miřič*, dále neživotné (*cídit/cidič, přímit/přimič*).

Poznámka: U substantiva *průvodič* se objevuje výjimečná alternace. Na samohláskové alternaci v prefixech (*pro-/prů-*) jsme u derivace deverbativ na *-č* nenarazili. Substantivum od slovesa *provádět* by mělo být tvořeno pravidelně a znít *prováděč* jako *prodavač, promítač, prohlížeč.* Alternace *pro/prů* je běžná (nikoliv bezpodmínečná) u dějových jmen tvořených bezafixálně (*průchod, průkaz, průvan, průjezd/projezd, průrva, průtah, průtok, ...* ale *provoz, prodej, projev, pronájem, ...*) a u substantiva *průvodce* (na rozdíl od *prodejce, pronájemce, provozce*). V korpusu jsme narazili na substantivum *průtokáč* tvořené univerbizací od *průtokový ohřívač* (deverbativum od slovesa *protékat* by asi znělo *protékač*).

.*áCt/.*a.eč	.*éCt/.*e.ač	.*áCt/.*a.ač*	.*íCt/.*0.eč**

*V korpusu byly nalezeny doklady dalších alternací KoS *krást/kradač* (.*áCt/.*a.ač), vokálem vloženým mezi kmen a sufix *-č* je v tomto případě *-a- (pletač)* KoS je ovšem *-á- (kladeč).*

** Zajímavý je doklad zánikové alternace KoS (*číst/čteč* - .*íCt/.*0.eč),

Poznámka: Deverbativum *okradač* pokládáme za derivát slovesa *okrádat* nikoli *okrást* vzhledem k tomu, že deriváty na *-č* jsou pravidelně tvořeny od imperfektiv (srv. výše). Jednou z výjimek doloženou též v korpusech je botanický termín *osladič.*

.*á.(e|ě)t/.* (a|á).(e|ě)č*	.*í.(e|ě)t/.* (i|í).(e|ě)č**

*K typu .*á.(e|ě)t/.* (a|á).(e|ě)č *sázet/s(a|á)zeč* můžeme na základě korpusů uvést substantiva jako *zaražeč, navažeč, rozvažeč, odhaněč, potapěč.*

**K typu .*í.(e|ě)t/.* (í|í).(e|ě)č *odklízet/odkl(i|í)zeč* jsme nalezli doklad *oviјet/oviјeč.*

.*á.at/.*a.ač*	.*í.at/.*ě.ač**	.*í.at/.*i.ač**	.*ou.at/.*u.ač**	.*ý.at/.*y.ač***

*K typu .*á.at/.*a.ač (.*á.at/.*a.(a|á)č) jsme nalezli např. *přespávat/přespavač, práskat/praskač, okrádat/okradač, dávat/davač,* dále např. jména prostředků *přehrávat/přehravač, rozeznávat/rozeznavač, prohledávat/prohledavač, nasávat/nasavač, vyorávat/vyoravač.*

**K typům .*í.at/.*ě.ač, .*í.at/.*i.ač a .*ou.at/.*u.ač jsme žádné další doklady nenalezli.

*** K typu .*ý.at/.*y.ač bylo nalezeno *umývat/umyvač.*

Poznámka: Substantivum *umyvač* je doloženo v PSJČ, ve slovníku automatického analyzátoru *ajka* i v SSJČ je pouze *umývač*, korpusy dokládají užívání substantiva *umyvač*.

Korpusy poskytují doklady pro zamyšlení nad alternacemi KoS z hlediska kodifikace dublet. Je jisté, že při malém počtu dokladů může jít pouze o překlepy (*potapěč, počitač, nahravač, ...*). Je ale také možné, že jde o důsledek rozkolísanosti úzu. Pokud budeme chtít formulovat přesná pravidla uvedených alternací, bude nejdříve nutné odpověď na otázku, z čeho by měla taková pravidla vyjít (jak reflektovat kodifikaci a normu, a to především v případech, kde korpusy ukazují možný rozpor).

Poznámka: Máme na mysli případy, kdy tištěné slovníky i slovníky morfologických analyzátorů zaznamenávají např. pouze deverbativum *sekáč,* nebo *zatykač* a korpusy dokládají též *sekač, zatýkač* atp.

Příklady deverbativ odvozených sufixem *-č* od sloves 3. třídy podle kmene přítomného tradičně řazených ke vzoru *krýt* jsou co do +/- alternací KoS poměrně těžko formálně popsatelné. Sledovali jsme proto jednotlivá slovesa a hledali jsme k nim příslušná deverbativa. Kromě alternujících případů jako jsou *mýt/myč,* existují názvy prostředků patřící do centra slovní zásoby, u nichž k alternacím nedochází (srv. *rýt/rýč* narozdíl od alternujícího *bít/bič*). V korpusech jsme navíc nalezli činitelská substantiva od slovesa *šít/šič*, *krýt/krýč* (sportovní profesionalismus – ten, kdo kryje), okazionalismy jako *žít/ žíč* (ten, kdo si žije), *pít/čajpíč* a *smát/smáč,* u nichž rovněž nedochází k alternacím KoS.

Vzhledem k vysokému počtu případů, kdy slova na *-áč* nejsou dverbativy, jsme pro potřeby automatické morfologické analýzy pátrali po dalších příkladech činitelských jmen na *-áč.* Seznamu deverbativ, u nichž se KmS *-á-* nekrátí (*hráč, spáč, rváč, štváč, zváč, žváč, dráč, práč, sráč, ...* i vlastní jméno *Tkáč, ...*), jsme doplnili o okazionalismus *psát/psáč* doložené v korpusech.

5. Závěr : formální pravidla derivace deverbativ na *-č*

Možnými kandidáty jmen činitelských na *-č* jsou slova, jejichž lemma (základní tvar) končí na *-č* a zároveň se jedná o maskulina životná (maskulina neživotná jsou naopak kandidáty na názvy prostředků, které jsou v některých případech tvořené paralelně např. *nosič*, v jiných bez paralely např. *klepáč/klepač* tímtéž sufixem).

Pokud má jít o deverbativum (jméno činitelské/prostředku), pak po formální stránce může po *-č* u jednotlivých tvarů následovat pouze uvedený řetězec znaků:

.*č(e|i|ů) nebo .*č(e|ů)m nebo .*čích nebo .*čov(i|é). Derivační pravidla mohou využít pravidla flexe zahrnutá v automatických analyzátorech (konkrétně lingvistickou bázi analyzátoru *ajka* srv. Osolsobě, 1996, Sedláček, 2004).

Před sufixem *-č* nemůže, má-li jít o deverbativum, předcházet konsonant (slova jako *pinč, Renč, ranč, pomeranč, punč, terč, ...* nemohou být deverbativa).

Před *-č* vždy předchází samohláska. Má-li jít o deverbativum, pak ji z lingvistického hlediska je možné interpretovat buď jako kmenovou samohlásku (KmS např.: *hlíd-a-t/hlíd-a-č*), nebo jako kořenovou samohlásku (KoS např.: *mý-t/my-č*), nebo jako vokál vložený mezi slovesný kmen (pouze u sloves 1. třídy podle kmene přítomného vzor *nést, péct*: *kradač, čteč, ...*) a sufix *-č* (v literatuře uváděnými i v korpusech a strojových slovnících doloženými výjimkami jsou substantiva *trubač, vozač, kropáč*). Pokud je samohláskou *-e,* pak při tvoření flektivních tvarů nikdy nedochází k zánikové alternaci (neplatí .**Ceč,>.*Cče :* substantiva skloňovaná jako *Bubeneč/Bubenče, Ledeč/Ledče, ...* nemohou být deverbativy na *-č, čteč/čteče* ano).

KmS může být pouze *-á-, -a-, -e-, -ě-, -i-, -í-*.

KoS může být libovolný vokál, nicméně se zdá, že v češtině nejsou doloženy případy, kdy by jím bylo *-o-, -ó-, -u-, -ů- (-ú-), -é-*. Mezi slovesy typu *krýt* není sloveso na *-(o|ó|ů|ú|é)t*. Slovesa na *-ou-t*, se sice vyskytují (*obout, zout, kout, plout, ...),* deverbativa na *-č* odvozená od těchto sloves nejsou ve zkoumaném materiálu doložena. Automatické přegenerovávání **kout/kouč* je třeba při aplikaci pravidel vyloučit.

V češtině jsme našli pouze doklady, kdy KoS bezprostředně předcházející před *-č* je *-á-, -a-, -e-, -ě-, -i-, -í-, -y-, -ý-*.

V češtině jsme našli pouze doklady, kdy vokál vložený mezi slovesný kmen (pouze u sloves 1. třídy podle kmene přítomného vzor *nést, péct* a výjimky *trubač, vozač, kropáč*) a sufix *-č* je *-a-, -á-, -e-*.

Pokud je substantivum na *-č* proprium, je třeba uvážit, zda má být popsáno v rámci popisu derivace apelativ. Vyloučení proprií má dobré (nejen praktické) důvody.

Pokud budeme nadále uvádět slovesné třídy a vzory, upozorňujeme na to, že je možné využít pravidel tvoření jednoduchých tvarů sloves od jednotlivých slovesných kmenů popsaných jako slovesné vzory (Osolsobě, 1996) a aplikovaných v automatickém morfologickém analyzátoru *ajka* (Sedláček, 2004). Algoritmický popis tvoření slovesných tvarů (Osolsobě, 1996) zachovává v základních rysech lingvistický přístup (Komárek, 1986), který ovšem na mnoha místech pro potřeby algoritmizace zjemňuje.

Automatická morfologická analýza, jejímž cílem by bylo nalézt dvojice sloveso/deverbativum na *-č*, by musela vyjít i z popisu doložených hláskových alternací KoS, pokud KoS nepředchází bezprostředně před *-č*. Na tomto místě pouze odkazujeme k tomu, co bylo řečeno výše (srv. odd. 2, 4).

Pokud před *-č* předchází *-í-*, pak deverbativa jsou pouze a) od sloves typu *umřít* (*dříč, tříč,* v korpusech např. kompozita *hlad(o|y)mříč),* b) od sloves typu *bdít (bdíč, mstíč, sníč)* a c) od sloves typu *krýt (žíč)*. U slovesa *mlít*, které některé gra-

matiky řadí ke vzoru *umřít,* je ovšem doloženo pouze činitelské *mleč*, nikoli *mlíč* (odvozeno ze substantiva *mlíčí* srv. *javor mlíč*). Aby se vyloučilo „přegenerovávání" je třeba vyloučit generování **bít/bíč (beech), *klít/klíč,* (*neklíč, centrklíč, multiklíč*), **mít/míč, *mlít/mlíč, (pumlíč), *sít/síč.* V analyzovaném materiálu nebyla doložena alternace KmS .*(e|ě)t/.*íč. Pravidlo, jehož aplikace by vedla k „přegenerovávání" **klet/klíč, *mlet/mlíč, *prostřednět/prostředníč* by nemělo existovat.

Pokud před *-č* předchází *-á*, pak může jít o desubstantivní názvy podle význačného rysu (typ *hlaváč*), deadjektivní názvy nositelů vlastností (typ *zelenáč*), desubstantivní názvy prostředků (typ *květináč*) aj (viz výše). Při aplikaci automatické analýzy může pomoci fakt, že k desubstantivům, deadjektivům, substantivům vzniklým univerbizací a krácením nemusí být doložena „fundující slovesa". Problematické mohou být případy, kdy by aplikace derivačních pravidel mohla vést k „přegenerovávání" (k příslušnému lemmatu na *-áč* by bylo nalezeno existující lemma slovesa na *-át/-at*). Aby se vyloučilo „přegenerovávání", je třeba vyloučit jako výjimky generování dvojic **plát/pláč, *tát/táč, *sekat(secondhand)/sekáč, *volat/voláč, *rubat/rubáč, *tutat/tutáč, zelenat/zelenáč, klepetat/klepetáč, krkat/krkáč.*

Zdá se, že na *-áč* mohou končit pouze deverbativa derivovaná od uzavřeného seznamu sloves. KmS/KoS *-á* se u nich nekrátí. Aplikaci pravidel při automatické analýze mohou ovšem komplikovat kompozita, k nimž neexistuje odpovídající slovesné kompozitum (**solohrát/sólohráč, *zvonkahrát/zvonkohráč, *sedmispát/sedmispáč, ...*).

Na *-ič* mohou kromě deverbativ, a to jak od sloves na *-it* (*rodit/rodič*), tak výjimečně od sloves na *-ít* (např. *šič, bič*) a *(e|ě)t* (např. *držič, prdič)*, končit též slova, která nejsou deverbativy např. apelativum (*carevič,* neživ. *sendvič*). Vysoké procento „nedeverbativ" představují propria, a to zejména jihoslovanská (*Karadžič, Klimovič, ...*). Při aplikaci automatické analýzy může pomoci fakt, že k nedeverbativům nemusí být doložena „fundující slovesa". Problematické mohou být případy, kdy by aplikace derivačních pravidel mohla vést k „přegenerovávání" (k příslušnému lemmatu na *-ič* by bylo nalezeno existující lemma slovesa na *-ít/-it*). Uvádíme „homonymii" doloženou ve sledovaném materiálu: **snoubit/snoubič.* Z potencionálně deverbativních proprií uvádíme substantivum *Dědič*, které automatický analyzátor použitý pro značkování korpusů ČNK omylem lemmatizuje jako apelativum lemma *dědič*. Tištěné slovníky (PSJČ, SSJČ, SSČ) uvádějí pouze substantivum *dědic* motivované substantivem *děd.* Mezi výjimky je třeba zařadit blokaci automatického generování dvojice **dojít/dojič.*

Pokud by při aplikaci automatické analýzy došlo k tomu, že k jednomu tvaru na *-ič* by se generovaly dva tvary, a to jeden na *-it* a druhý na *-(e|ě)t*, pak by mělo platit pravidlo, že má přednost tvar na *-it*. (Je logické, že činitelská jména se tvoří od faktitiv (*tupit*), nikoliv od inchoativ (*tupět*) viz výše).

Na *-ač* mohou kromě deverbativ končit nedeverbativa např. apelativa *(čuvač, knírač)*. K desubstantivům, deadjektivům, substantivům vzniklým univerbizací a krácením nemusí být a v analyzovaném materiálu nejsou doložena „fundující

slovesa“. Aby se vyloučilo „přegenerovávání“ je třeba vyloučit generování **bát/bač* (tvary *bače, bačů* substantiva *bača*).

Na *-yč/-ýč* mohou kromě deverbativ *(mýt/myč, krýt/krýč, rýt/rýč)* končit i nedeverbativa, a to především propria na *-yč* (*Fryč, Maděryč, ...*). Při aplikaci automatické analýzy může pomoci fakt, že k nedeverbativům nemusí být doložena „fundující slovesa“. Problematické mohou být případy, kdy by aplikace derivačních pravidel mohla vést k „přegenerovávání“ (k příslušnému lemmatu na *-yč/-ýč* by bylo nalezeno existující lemma slovesa na *-yt/-ýt*). Aby se vyloučilo „přegenerovávání“ je třeba vyloučit jako výjimky generování dvojice **týt/tyč*.

Na *-eč/-ěč* mohou kromě deverbativ (*sazeč, pohaněč, mleč*) končit též propria (*Peč, Bubeneč, Ledeč, ...*), která mohou být v korpusech chybně lemmatizována i značkována. Při aplikaci automatické analýzy může pomoci fakt, že k nedeverbativům nemusí být doložena „fundující slovesa“. Aby se vyloučilo „přegenerovávání“ je třeba jako výjimky vyloučit generování dvojic **klít/kleč, *lít/leč, *mít/meč, *pít/Peč, *sít/seč, *tít/teč, *vít/več*. Při aplikaci je třeba dát pozor na překlepy či pokusy zaznamenat vady výslovnosti doložené v korpusech (*šeč, žeč, čeč*). I ony by při analýze neoznačkovaného textu mohly být důvodem „přegenerovávání“ a vytváření nesmyslných dvojic *šít*/**šeč, žít/*žeč, čít/*čeč*.

Automatická morfologická analýza založená na výše uvedených pravidlech by u nelemmatizovaných a neoznačkovaných tvarů nebyla schopna diferencovat rozdíl životná/neživotná maskulina (jména činitelská/prostředků) u tvarů, jejichž koncovky jsou homonymní, tj. všechny kromě .*čov(i|é) a homonymie *snoubič(<snmoubit/snowboard), sekáč(<sekat/secondhand)*.

5.1 Shrnutí

Cílem této studie bylo ověřit, jak lze využít nástroje *Deriv* a jazykových korpusů k přesnější formulaci pravidel derivace jednoho slovotvorného typu (deverbativ na *-č,* především jmen činitelských). Na základě analýzy materiálu získaného ze slovníku morfologického analyzátoru *ajka* a z korpusů ČNK navrhujeme formální pravidla derivace sledovaného typu (srv. odd. 5).

Pravidla, která jsme na základě výsledků testování automatického nástroje *Deriv* a analýzy dat doložených v korpusech ČNK navrhli, by mohla být využita v aplikacích např. při přípravě podkladů pro derivační slovník češtiny, ale také při tvorbě automatických guesserů založených na pravidlech opřených o lingvistické zákonitosti. Materiál získaný z korpusů lze rovněž využít v lexikografické praxi.

LITERATURA

DOKULIL, M.: Tvoření slov v češtině. Praha : Československá akademie věd 1962.

DOKULIL, M., KOMÁREK, M. a kol.: Mluvnice češtiny I, II., Praha : Academia 1986.

HAJIČ J.: Desambiguation of Rich Inflection (Computational Morphology of Czech). Praha : Karolinum, Charles University Press 2004.

HAVRÁNEK, B. a kol.: Slovník spisovného jazyka českého (SSJČ). Praha : Academia 1989.

HUJER, O., SMETÁNKA, E., WEINGART, M., HAVRÁNEK, B., ŠMILAUER, V., ZÍSKAL, A. (red.): Příruční slovník jazyka českého (PSJČ). Praha : Státní nakladatelství 1935-1957.
FILIPEC, J. a kol.: Slovník spisovné češtiny pro školu a veřejnost (SSČ). Praha : Academia 2005.
HLAVÁČKOVÁ, D., PALA, K.: Computer Processing Derivational Relations in Czech. In Computer Treatment of Slavic and East European Languages. Bratislava : Slovak National Corpus, L. Štúr Institute of Linguistics, Slovak Academy of Sciences 2007, s. 198-208.
HLAVÁČOVÁ, J.: Morphological Guesser of Czech Words. In Matoušek, V. (ed.): Text, Speech and Dialogue. Berlin : Springer-Verlag 2001, s. 70-75.
KARLÍK, P., NEKULA, M., RUSÍNOVÁ, Z.: Příruční mluvnice češtiny. Praha : NLN 1995.
KLÍMOVÁ, J., ŠTÍCHA, F.: K možnostem a mezím sufixální derivace substantiv. In: Štícha, F. (ed): Možnosti a meze české gramatiky. Praha : Academia 2006, s. 127-138.
KRÁLÍK J., Těšitelová M.: Retrográdní slovník současné češtiny. Praha: Academia 1986.
OSOLSOBĚ, K.: Algoritmický popis české morfologie a strojový slovník češtiny. Brno : FF MU, (disert. práce) 1996.
OSOLSOBĚ K., Pala, K., Sedláček, R., Veber, M.: A Procedure for Word Derivational Processes Concerning Lexicon Extension in Highly Inflected Languages, In: Proceedings of the Third International Conference on Language Resources and Evaluation, LREC, Las Palmas de Gran Canaria : ELRA 2002, s. 1254-1259.
RYCHLÝ, P.: Korpusové manažery a jejich efektivní implementace. Brno : FI MU (disertační práce) 2000.
RYCHLÝ, P.: Bonito – grafické uživatelské rozhraní systému Manatee, Verze 1.49. (1998-2003). Dostupné z http://ucnk.ff.cuni.cz/bonito/
SEDLÁČEK, R.: Morphematic analyser for Czech. Brno : FI MU, (disert. práce) 2004.
SLAVÍČKOVÁ, E.: Retrográdní morfematický slovník češtiny. Praha : Academia 1975.
ŠEVEČEK, P.: Morfologický analyzátor (lemmatizátor) Lemma, program v jazyce C. Brno : FI MU (disert. práce) 1996.
ŠMILAUER, V.: Novočeské tvoření slov. Praha : Státní pedagogické nakladatelství 1971.
ŠMILAUER, V.: Nauka o českém jazyku. Praha : Státní pedagogické nakladatelství 1972.
Český národní korpus - SYN2000/SYN2005. Ústav Českého národního korpusu FF UK, Praha 2000. Dostupný z WWW: <http://ucnk.ff.cuni.cz>. (http://ucnk.ff.cuni.cz/bonito/) *Deriv* http://nlp.fi.muni.cz/projekty/ajka/cjbb85/

ALGORITHM FOR DERIVATION OF DEVERBATIVS FORMED BY *-Č*

A new tool (http://nlp.fi.muni.cz/projekty/ajka/cjbb85/) for the computer processing of the derivational relations (*Deriv*) and its proper usage is tested in the paper (1, 2). Occurence of potentional deverbativs on *-č* is pursued in Czech Corpora (SYN2000, SYN2005, SYN2006PUB) (3), and the material for further analysis is extracted both from corpus and from the machine dictionary (4). Either sources serve for delimination of the word formation's formal rules and of the list of possible overgenerations and homonyms (5). The formal rules could be employed in the automatic developing of the derivational dictionary of Czech and for the creation of morphological guesser based on linguistical regularities.

Klára Osolsobě
Ústav českého jazyka
Filozofická fakulta Masarykovy univerzity
Arna Nováka 1
602 00 Brno
e-mail: osolsobe@phil.muni.cz