Propria (příjmení na -*č*) - problém automatické morfologické analýzy

Klára Osolsobě

osolsosobe@phil.muni.cz

„Co pořád máte s tím datlem?“

(Rabbi Jacob)

Zdrojem jazykové komiky na straně jedné a kamenem úrazu v oblasti aplikací NLP (natural language processing - počítačové zpracování přirozeného jazyka) na straně druhé jsou nejrůznější případy víceznačnosti, které se v jazyce běžně vyskytují. Jednou ze zhusta citovaných „homonymií“ je případ, kdy slovotvorný modul (počítačový program, který má za úkol generovat/analyzovat utvořená slova) mylně přiřadí substantivum *datel* ke slovesu *dát.* Pokud by ovšem příslušná aplikace (slovotvorný modul) byla propojena s automatickým morfologickým analyzátorem (modulem tvarotvorným), mělo by být nemožné k uvedenému výsledku dojít, neboť všechna deverbativa (činitelská jména na *–tel*) se v nové češtině 1) skloňují podle měkkého vzoru (*muž/učitel*) a 2) při flexi u nich nedochází k samohláskové alternaci (při tvoření tvarů s nenulovou koncovkou nemizí *-e-*, které je součástí sufixu *-tel*). Obě tato jednoduchá pravidla jsou v případě substantiva *datel* porušena. V této studii opustíme rovinu žertovnou a pokusíme se ukázat, jakým omylům je třeba předcházet při pokusech o aplikaci automatické syntézy/analýzy derivace slov v češtině. Naší pozornosti se budou v zájmu jubilantky těšit vlastní jména, a to příjmení na *-č*.

Poznámka: Jména činitelská tvořená sufixem *–tel* se v češtině jako příjmení téměř nevyskytují. Příjmení nevznikla ani z přídomků *(Vilém Dobyvatel)*, či přezdívek *(Jan Křtitel, Adam Stvořitel)*. Situace je jiná u derivátů (nejen deverbativ) na *–č.* Ty se totiž od nejstarších dob jako příjmí, posléze příjmení objevují.

Automatická morfologická analýza proprií

Jedním z problémů automatické morfologické analýzy užívané v nejrůznějších aplikacích k morfologickému značkování jazykových korpusů je nutná omezenost strojových slovníků, nad nimiž příslušné programy operují, na straně jedné a neomezené bohatství slovních tvarů v textech, které tvoří jazykové korpusy, na straně druhé. Automatické nástroje sloužící k automatickému značkování korpusů se navíc nebudují pro jeden jediný korpus s vymezeným a omezeným rozsahem slovních tvarů a lemmat. Mezi jednotkami, které se v textu náhodně vyskytují s malou frekvencí, se často objevují mj. též propria. Zpracování proprií v rámci automatické analýzy naráží na řadu problémů. Prvním je otázka, jak řešit tokenizaci tj. rozčlenění textu na jednotky (pozice), které budou předmětem další analýzy.

Tokenizace vlastních jmen souvisí s velmi rozsáhlou oblastí automatického zpracování tzv. MWD (multi words expressions - víceslovných výrazů) jako např. *Ho či min, F. X. Šalda, Ludwig wan Beethoven, Nové Město nad Metují, ...*, kterou ovšem v rámci této studie ponecháme stranou. Druhou otázkou je otázka lemmatizace. Pokud je ve slovníku automatického morfologického analyzátoru interpretovaný tvar doložen pouze jako proprium, pak lemma začíná velkým písmenem, takže tvaru *Pavlovi* přiřadí automatická morfologická analýza lemma *Pavel.* Pokud je interpretovaný tvar doložen i jako apelativum (homonyma), pak lemma začíná buď písmenem malým nebo velkým. Výsledky disambiguace tj. přiřazení jednoznačné interpretace v případě, že automatická analýza nabízí více řešení (např.: <***Luka***> *v zimě, ano v zimě jsou* <***luka*** > *jiná.* <***Luka*** > *zvou na noční lyžování.* <***Luka*** > *Petrovič funěl jako lokomotiva. Já jsem* <***dán***> *symbiózou karikatury a ilustrace. Já jsem* <***Dán>*** *a kašlu na politiku*.), nejsou příliš zdařilé. Důvodem je snad i to, že využívání lingvisticky motivovaných intuicí (např.: pokud řetězec uvnitř věty – formálně tehdy, když nenásleduje po znaku „.“ nebo „!“ nebo „?“ nebo „:“– začíná velkým písmenem a zároveň je ve slovníku interpretován jako slovo, jehož lemma může začínat velkým písmenem, pak je třeba dát této interpretaci přednost) naráží na řadu problémů. Propria, a to především příjmení, výrazně narušují uvedené pravidlo tím, že se nezřídka vyskytují po znaku „.“ (tečka), který stojí za iniciálou křestního jména (např. *T. Mazáč, F. Topič, ...*). Pro automatickou analýzu, která neumí rozlišit tečku za zkratkou a tečku jako interpunkční znak signalizující konec věty, je pak příjmení stojící za iniciálou křestního jména nerozpoznatelné bez dalších pravidel od případu, kdy se slovo (libovolné) na začátku věty píše s velkým počátečním písmenem. Při lemmatizaci proprií je tedy poměrně nesnadně řešitelný problém lemmatizace proprií, která v jazyce figurují (mouhou figurovat) jako apelitava. Sem patří i celá řada příjmení.

V naší studii se ovšem budeme věnovat pouze jedné z otázek, která souvisí s homonymií apelativum/proprium. Na základě analýzy proprií doložených v korpusech ČNK, konkrétně příjmení na *–č*, se budeme snažit doplnit seznamy výjimek, které mohou pomoci zabránit „přegenerovávání“ při aplikacích pravidel automatické analýzy derverbativ na *–č* (činitelská/prostředků). V tomto článku navazujeme na formální popis derivace deverbativ na *–č* (Osolsobě 2008) a věnujeme se propriím, která v citované studii zůstala stranou našeho zájmu.

Poznámka: V analýze budeme vycházet z materiálu tří korpusů psaného jazyka. SYN2000 (100 milionů slovních tvarů), SYN2005 (100 milionů slovních tvarů) jsou vyvážené (zahrnují texty v udaném percentuálním poměru zastupujících všechny funkční styly). Korpus SYN2006PUB (300 milionů slovních tvarů) zahrnuje výhradně publicistické texty.

Automatická morfologická analýza a problém „přegenerovávání“

Jedním z podstatných rysů aplikací automatické analýzy je tzv. přegerovávání, tj. gnerování tvarů, které jsou z hlediska systému automatické analýzy „správné“, ale z hlediska úzu jsou buď periferní, nebo nejsou doložené, popřípadě jsou homonymní s tvary, které nejsou utvořeny podle příslušných pravidel, ale odpovídají pravidlům jiným (např. podle jednoho z pravidel pro generování deverbativ na *–č* se správně generují dvojice *bít/bič*, *mýt/myč, rýt/rýč, šít/šič* a dokonce i *krýt/krýč, žít/žíč*, nesprávně pak dvojice **mít/míč, *mlít/mlíč, *týt/tyč*, podle jiného z pravidel navržených pro automatické generování deverbativ na *–č* se správně generují dvojice *kopat/kopáč, mazat/mazáč, orat/oráč* a nesprávně dvojice **krkat/krkáč, *tutat/tutáč, *volat/voláč, *zelenat/zelenáč*).

Zabránit přegenerování je jedním z cílů s nimiž se automatická morfologická analýza musí vyrovnávat. Jedním z prostředků omezení přegenerovávání je pokud možno co nejrozsáhlejší přehled o nejrůznějších výjimkách z formálních pravidel. Korpusy poskytují díky svému rozsahu (stamilióny slovních tvarů) a povaze (empirická data) celou řadu informací, které si donedávna nebylo dost dobře možné představit. Jejich předností je mimo jiné i to, že se v nich objevují slova nezachycená v dosavadních popisech jazyka (především ve slovnících), a to ve svém přirozeném užití, které je možno zkoumat.

Příjmení na *–č* v korpusech ČNK

V následující tabulce uvádíme přehled počtu lemmat začínajících velkým písmenem (tvarů analyzovaných automatickou morfologickou analýzou jako propria), zakončených na *–č*, jimž automatická morfologická analýza přiřadila značku maskulinum životné (*NNM.**). Rozdělujeme je (řádky) podle toho, zda před *–č* předchází konsonant (C), nebo vokál.

[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“** & tag=“NNM.*“]

	SYN2000	SYN2005	SYN2006PUB
Cč (*Renč*)	8	62	18
ač (*Hubač/Zuvač*)	33	128	52
áč (*Hlaváč/Mazáč*)	37	78	53
eč/ěč (*Pec/Pěč*)	0	15	2
éč (*Péč*)	0	0	1
ič (*Topič/Frič*)	548	918	911
íč (*Síč/Píč*)	1	7	3

oč (*Kaloč*)	2	21	5
uč (*Bezruč/Rouč*)	3	15	3
úč (*Balúč*)	1	1	1
yč (*Fryč/Nyč*)	3	45	4
ýč (*Nýč*)	1	2	2
ostatní	0	15	0
CELKEM	637	1305	1055

Poznámka: Při značkování korpusu SYN2005 byly použity tzv. guessery (hadače), tj. automatické programy zaměřené na „uhadování“ informace u slov, kterým automatický analyzátor nepřiřadí ani jednu interpretaci. Tyto automatické nástroje bývají založeny na různých přístupech (lingvistických, statistických, kombinovaných) (srv. Hlaváčová 2001). Používáním guesseru si lze vysvětlit nárůst počtu lemmat, která vyhovují zadaným požadavkům (přiřazení značky indikující maskulinum životné) v korpusu SYN2005 oproti korpusům SYN2000, SZN2006PUB, při jehož značkování guesserů použito nebylo. Mnoho tvarů je ovšem analyzováno chybně, a to jak na úrovni lemmatu, tak na úrovni morfologické značky. Větší počet lemmat u korpusu SYN2006PUB lze přičíst na vrub jeho rozsahu (300 milionů tvarů) i složení (publicistické texty).

Příjmení na *–č* s potenciální slovesnou motivací v korpusech ČNK

Apelativa (tedy i deverbativa na *-č*) mohou fungovat jako propria (srv. např. příjmení typu *Mazáč*, *Topič, ...*). Pro průzkum příjmení na *-č*, která mohou být odvozena od sloves jsme užili materiál získaný ze synchronních korpusů ČNK. Zvolili jsme následující postup.

Vyhledali jsme všechna lemmata začínající velkým písmenem a končící na *–ač, -áč, -eč, -ěč, -ič, -íč, -yč, -ýč*, mající značku substantiva, maskulina životná (*NNM.**). Tento postup vychází z analýzy jazykových dat získaných z rozsáhlého strojového slovníku českých kmenů (Osolsobě 1996) a z korpusů ČNK, na jehož základě se ukázalo, že lemmata substantiv, která mohou být v češtině deverbativy na *–č*, musí splňovat dvě formální podmínky: 1) před *–č* nesmí předcházet konsonat a 2) vokál, který předchází před *–č* není *–é-, -o-, -ó-, -u-, -ů-, (-ú-)* (Osolsobě 2008).

[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).*(**
(a|á|e|ě|i|í|y|ý)č“ & tag="NNM.*"]

Seznamy lemmat jsme podrobili ruční analýze a vytřídili jsme propria, ke kterým by bylo možné na základě formálních pravidel derivací sloveso - činitelské jméno na *–č* (Osolsobě 2008) generovat automaticky existující česká slovesa. V následující tabulce uvádíme přehled počtu lemmat, která jsou po formální stránce kandidáty na propria od původu deverbativa na *–č* (KDČ).

KDČ	SYN2000	SYN2005	SYN2006PUB
ač (*Zuvač*)	12	18	20

áč (*Mazáč*)	9	15	23
ěč (*Pěč*)	0	0	1
ič (*Topič*)	7	14	22
íč (*Síč*)	0	2	2
yč (*Nyč*)	0	2	1
ýč (*Nýč*)	0	2	1
CELKEM	28*	53**	70***

* *Bukač/502, Machač/176, Tykač/153, Kopač/79, Salač/68, Piskač/56, Pleskač/25, Drač/10, Kovač/7, Tleskač/6, Takač/6, Dupač/2, Kováč/2269, Řezáč/260, Tkáč/117, Klepáč/79, Takáč/26, Mazáč/19, Klapáč/14, Klofáč/9, Tokáč/1, Čič/120, Zvonič/81, Radič/22, Bulič/7, Prudič/3, Hajič/2, Banič/1.*

** *Tykač/48, Kopač/46, Bukač/33, Salač/32, Machač/32, Piskač/21, Kovač/20, Dupač/20, Klapač/10, Pleskač/9, Trsač/4, Šťourač/3, Kukač/2, Sochač/1, Drač/1, Tlachač/1, Kutač/1, Lupač/1, Kováč/163, Řezáč/78, Tkáč/63, Klepáč/22, Takáč/19, Mazáč/16, Klofáč/8, Kudláč/5, Konáč/3, Klapáč/3, Lupáč/1, Kutáč/1, Kropáč/1, Kakáč/2, Běláč/1, Čič/10, Zvonič/9, Mastič/4, Radič/4, Pilič/3, Mič/3, Bilič/3, Banič/2, Sič/2, Bajič/2, Minič/2, Mršič/2, Prudič/1, Dymič/1, Píč/17, Síč/10, Nyč/18, Byč/2, Nýč/10, Týč/1.*

*** *Bukač/1861, Tykač/608, Machač/522, Kopač/324, Salač/265, Piskač/147, Kovač/92, Pleskač/61, Lupač/44, Klapač/29, Kukač/28, Drač/20, Rachač/14, Hupač/13, Takač/12, Dupač/12, Tlachač/10, Buchač/7, Bryndač/7, Kutač/7, Řezáč/1120, Tkáč/288, Klepáč/279, Takáč/181, Klofáč/142, Mazáč/134, Klapáč/89, Kutáč/45, Kudláč/18, Lupáč/10, Hrabáč/10, Konáč/4, Racháč/1, Čič/302, Zvonič/178, Pilič/94, Bilič/83, Radič/52, Matič/22, Banič/16, Jevič/16, Kudlič/16, Hajič/15, Bulič/11, Prudič/10, Mastič/8, Rosič/7, Mučič/7, Čelič/5, Mamič/5, Manič/5, Medič/4, Trudič/2, Spasič/2, Musič/1, Dražič/1, Síč/35, Píč/15, Nyč/193, Nýč/80, Pěč/3.*

Při nalezení jakékoli jednotky v korpusu je možné sledovat její kontext. Kontext vypovídá a napovídá mnohé o významu (motivaci) jednotek i o jejich užití. Situace je ovšem komplikovanější u proprií, jejichž motivace je mnohdy hůře odhalitelná, a na rozdíl od motivace apelativ nebývá možné ji uhádnout z kontextu. Proto jsme se opírali především o příslušnou onomastickou literaturu, v níž jsme nalezli výklad většiny sledovaných příjmení (Beneš 1962, 1970, Moldánová 2004).

V uvedených seznamech (číslice znamenají frekvenci lemmatu v příslušném korpusu) se na první pohled vyskytují příklady „přegenerovávání“ automatické morfologické analýzy, která přiřadí podle vzoru *vázat/vazač* proprium *Machač* k verbu *máchat*, zatímco pro *Macháč (*machat)* pravidlo nenalezne. Obě substantiva jsou tvořena od antroponyma *Mach* z osobního jména Mathias (Beneš 1962, Pleskalová 1998: 61).

Nejpravděpodobnější se zdá být motivace slovesným základem u jmen na *-ač/-áč.*

Slovníky uvádějí apelativa neživotná maskulina substantiva *dráč, kakáč, klapač, klapáč, klepač, klepáč, klofáč, tykač* (PSJČ) a životná maskulina *bukač, piskač, kutač* (SSJČ)*, pleskač* (PSJČ), *tleskač, tkáč* (SSJČ i PSJČ), dále feminina *dupačky, šťouračka, dračka (jít na dračku), tlachačka.* Alternace kořenové samohlásky u apelativních deverbativ jsou dosti rozkolísané, slovníky připouštějí v řadě případů dublety. Dublety se objevují i u proprií (srv. apelativum *dráč* a propria *Dráč* (Beneš 1962: 290) i *Drač* doložené ve sledovaných korpusech.

U substantiv na *–ič* jsou doložena apelativa životná maskulina *hajič, radič* (SSJČ)*, zvonič* (PSJČ, SSJČ).

Mezi příjmeními na *-ač/-áč* jsme v příslušné onomastické literatuře (Beneš 1962: 285 nn., Moldánová 2004) našli potvrzenu slovesnou motivaci u příjmení *Bryndač, Dráč* (nikoli *Drač), Dupač, Hupač, Kakač/Kakáč*, *Klapač, Klepáč, Klofáč, Kukač, Lupáč/Lupač*, *Mazač/Mazáč Pleskač*, *Štourač/Šťourač, Tlachač, Tykač,* dále *Bukač, Hrabáč, Kováč, Kropáč, Kutač, Řezáč, Tkáč* (ibid.: 2004). Beneš řadí některá z nich do jednotlivých skupin jako např. příjmení z názvů čeledi a podobných stálých pracovníků, příjmením z názvů výrobců textilního zboží a oděvníků atd., slovesná motivace je přípustná.

U propria *Piskač* uvádí Beneš (1962: 225) motivaci substantivem *pysk* (příjmení z názvů částí obličeje a hlavy), zatímco Moldánová (2004: 140) se přiklání k motivaci slovesem (srv. apelativum *pískač*, řidč. *piskač* synonymní k *pištec*).

Příjmení *Buchač/Bucháč* jsme v příslušné literatuře nenalezli, našli jsme ovšem příjmení *Bucháček,* které je interpretováno buď jako příjmení od slovesného základu (Beneš 1962: 290), nebo jako příjmení od osobního jména *Budislav* (Moldánová 2004: 35).

Adjektivní motivaci mají dle literatury jména *Běláč* a *Kudláč.* Automaticky generované dvojice **bělat/Běláč*, **kudlat/Kudláč* jsou případem hledaného přegenerovávání.

Příjmení *Racháč, Salač* pocházejí z osobních jmen *Rach* (Moldánová 2004: 152), *Salamon* (ibid.: 164). Automaticky generované dvojice **rachat/Racháč* **sálat/Salač* jsou dalšími případy hledaného přegenerovávání.

Interpretaci příjmení *Konáč, Sochač, Trsáč, Tokáč* jsme nenalezli v uvedené literatuře. Domníváme se, že příjmení *Konáč* je téhož původu jako *Konečný* (substantivní motivace). Se substantivní motivací se běžně setkáváme u apelativ na *–áč* (*kafáč, květináč, malináč, ...*). Ačkoliv jsme nenalezli výklad příjmení *Trsač, Sochač,* ale pouze *Trs, Trsek* (příjmení z názvů rostlin (pícnin) a *Soch, Socha, Sochorek* (příjmeními z názvů dřevěných nástrojů) (Beneš 1962), jeví se nám přijatelnější motivace substantivy (*socha, trs*) než interpretace nabízená automatickým generováním **sochat/Socháč, *trsat/Trsáč*. U příjmení *Tokáč* váháme. V příslušné literatuře jsou doložena příjmení *Tocháč* od osobních jmen *Tobiáš/Tomáš* a *Tokan* od slovesa *tokat* (Moldánová 2004).

Příjmení *Takač/Takáč* pochází z maďarštiny (*Takacz*), kde znamená *tkadlec* (Moldánová 2004: 194). Automatické generování dvojice **takat/Takáč* je opět příkladem přegenerovávání.

Nejvíce je přegenerovávání patrné u proprií na *–ič* vesměs jihoslovanských vlastních jmen. Slovesnou motivaci potvrzuje příslušná literatura u jmen *Hajič, Radič, Sič, Zvonič.*

U příjmení *Prudič* (v korpusech je doloženo i apelativum *prudič*) se uvádí motivace adjektivem *prudký* (Moldánová 2004).

Přestože slovníky i korpusy dokládají slovesa *bájit, bánit, bílit, bulit, čelit, čít, dražit, dýmit, jevit, kudlit, mámit, manit, matit, mastit, medit, mít, mršit, mučit, musit, pílit, radit, rosit, sít, spasit, trudit,* šlo by v případě automatického generování dvojic sloveso/substantivum na *–ič* doložené výše uvedenými proprii o příklad přegenerování automatické analýzy.

Z lingvistického hlediska je kromě toho, že se jedná o propria (která se chovají v rámci systému jinak než apelativa), možné vyjít z faktu, že deverbativa na *–č* se tvoří v naprosté většině případů od imperfektiv, což by ovšem mohlo pomoci vyloučit pouze chybné automatické generování dvojice **spasit/Spasič*.

V literatuře jsme nenalezli příjmení *Čič.* Beneš (1962: 163) řadí proprium *Čičová* do skupiny příjmení derivovaných od místních jmen (*Čičovice*), uvádí ovšem též příjmení z názvů savců (*kočička*) *Čičalka, Čičatka* (ibid.: 190). Pleskalová (1998: 65) uvádí *Čáč/Čieč* od vlastního jména *Čáslav.*

U jména *Bulič* by mohla připadat v úvahu příbuznost s příjmeními z názvů savců (Beneš 1962: 198) *Bulík, Bulíček*. U jména *Dražič* není pravděpodobná příbuznost s příjmeními ze sloves typu prosit (ibid.: 272) *Dražil*, či deverbativem (ibid.: 289) *Dražka*. Příhodnější se jeví příbuznost se jmény z obyvatelských názvů (ibid.: 167) (*Dražan*), či zdrobnělinami z osobních křestních jmen (ibid.: 109, 118, 123) *Dražek, Dražík, Dražka*.

Pro výklad dalších proprií na *–ič* doložených v korpusech se nám nepodařilo najít v odborné literatuře žádnou oporu. Dle kontextu (křestních jmen, rozvíjejících adjektiv „*chorvatský, jugoslávský, jihoslovanský, ...*“) jde o jihoslovanská příjmení.

U poměrně řídce doložených příjmení na *-íč, -yč, -ýč, -ěč*, se setkáváme především s motivacemi vlastními jmény. Jediný případ deverbativa je příjmení *Síč* (Moldánová 2004). V korpusech je sice doloženo apelativní kompozitum se zadním členem deverbativem na *–č čajpíč,* interpretace příjmení *Píč, Nýč, Týč, Byč, Pěč* jakožto deverbativ od sloves *pít, nýt, týt, být, pět,* nabízená automatickou analýzou, se nepotvrdila. Příjmení *Píč, Nýč, Týč* pocházející z německých příjmení (Beneš 1962) a jsou motivována osobními jmény *Petr, Nicolaus, Dietrich* (Moldánová 2004). U *Týč* a *Pěč* se uvádí motivace adjektivy *tiutch – deutsch* a *pěkný* (ibid.: 2004). V literatuře uvedené níže jsme nenašli výklad příjmení *Byč.*

Příjmení na *-č* odvozená od sloves v korpusech ČNK

Z celkového počtu kandidátů na propria od původu deverbativa na *–č* (KDČ) jsme v literatuře našli analyzovány tyto doklady proprií, které jsou od původu deverbativa na *–č* (DČ) pouze u některých kandidátů uvedených výše. Počty DČ uvádíme v tabulce.

	KDČ	DČ	přegenerovávání
SYN2000	28	8+6+2=17	11
SYN2005	53	13+10+3+1=27	26
SYN2006PUB	70	15+9+2=27	43
CELKEM různých	24+18+29+7=77	17+12+4+1=34	43

?Buchač, Bukač, Bryndač, Drač, Dupač, Hupač, Klapač, Kopač, Kovač, Kukač, Kutač, Lupač, *Machač, Piskač****, Pleskač,*** *Rachač, Salač, Sochač****, Šťourač,*** *Takač,* ***Tlachač, Tleskač, Tykač,*** *Trsač, Běláč,* ***Hrabáč, Kakáč, Klapáč, Klepáč, Klofáč,*** *Konáč,* ***Kováč, Kropáč,*** *Kudláč,* ***Kutáč, Lupáč, Mazáč,*** *Racháč,* ***Řezáč,*** *Takáč,* ***Tkáč,*** *?Tokáč,*
Bajič, Banič, Bilič, Bulič, , Čelič, Čič, Dražič, Dymič, ***Hajič,*** *Jevič, Kudlič, Mamič, Manič, Medič, Mastič, Matič, Mič, Minič, Mršič, Mučič, Musič, Pilič, Prudič,* ***Radič,*** *Rosič,* ***Sič,*** *Spasič, Trudič,* ***Zvonič,***
?Byč, Nyč, Nýč, Pěč, Píč, ***Síč****, Týč.*

Uvedené počty (v 56 % případů dochází k přegenerovávání) dokládají, že lingvistická analýza proprií nalezených v korpusech může být relevantním příspěvkem pro doplnění seznamů možných případů přegenerovávání v aplikacích automatického generování/analýzy slovotvorných vztahů.

Příjmení z deverbativních apelativ na *–č* v korpusech ČNK mylně analyzovaná jako apelativa

Jak bylo řečeno výše, jsou případy, výskytu apelativ v roli proprií špatně podchytitelné automatickou morfologickou analýzou. V důsledku toho je i řada propriálních užití v korpusech špatně analyzována na úrovni lemmatu. Například substantivum *Topič* je v následujících případech (a) ... *František <Topič/topič>, nakladatel a knihkupec* ... (b) *Jde o povídku <Topič/topič>, tvořící samostatnou kapitolu Ameriky* ... (c) ... *současná výstava v pražské galerii <Topič/topič>* ... lemmatizováno jako apelativum (/*topič*), tedy stejně jako případ (d) *<**Topič/topič**> kulturního domu se při požáru přiotrávil*. V naší studii jsem tyto poměrně složitě zjistitelné případy ponechali stranou.

V průběhu práce na přípravě popisu pravidel pro automatické generování deverbativ na *–č* (Osolsobě 2008) jsme nicméně na řadu případů náhodně narazili (např. *Barvič, Bělič, *Dědič, Fukač, *Hadač, Kopáč, Kulič, Lamač, Lepič, Mazač, Palič, Rubač, Topič, Voráč, Zuvač, ...*).

Poznámka: Tyto případy je možné systematičtěji vyhledávat v korpusech, a to tak, že vyhledáme všechna potencionální deverbativa na *–č*, která nemají lemma s velkým písmenem a zároveň tvar v korpusu (word) velkým písmenem začíná, přičemž podle značky jde o maskulina životná:

[word="**(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).*(a|á|e|ě|i|í|y|ý)č**" & 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).*(a|á|e|ě|i|í|y|ý)č**" & tag="NNM.*"]

a pomocí N-filtru odstraníme případy, kdy vyhledané slovo následuje po interpunkčním znaku „.“, „!“, „?“, tedy [tag="Z.*"].

Pro potřeby automatické analýzy byly zajímavé především případy, u nichž dochází při derivaci k hláskovým alternacím kořenového vokálu (*fukač, lamač, palič, zuvač*). Sporná je interpretace propria *Dědič* (mylně analyzovaného jako apelativum). Beneš (1962) uvádí propria *Dědic, Dědič* mezi příjmeními z názvů příslušníků selského poddaného lidu a svobodníků (ibid.: 211 cit.: „dědici byli původně svobodní a svůj majetek dědili ...“). Moldánová (2004: 41) uvádí u příjmení *Dědic, Dědič* motivaci substantivem *děd.* Za homonymní k deverbativnímu apelativu *hadač* (< *hádat*) lze pokládá proprium *Hadač* řazené k příjmením z částí slov (Beneš 1962: 76).

Závěr

Ačkoliv se rozsáhlé jazykové korpusy od doby vzniku korpusové lingvistiky budují především pro potřeby lexikografie, mohou sloužit a slouží i pro bádání v řadě dalších lingvistických oborů. V tomto příspěvku jsme se snažili demonstrovat, jak mohou korpusy přispět při zpracování lingvistických podkladů použitelných v oblasti NLP. Na příkladu analýzy vlastních jmen (maskulin životných na *–č*) jsme ukázali, na jaká úskalí mohou narazit pokusy vytvořit automatický modul derivace konkrétně českých deverbativ na *-č*. Výsledkem je rozsáhlý seznam možných případů přegenerovávání, kterých by se mohl automatický nástroj dopustit v případě, že by nepracoval dostatečně s rozdílem apelativum/proprium.

Propria (Family Names on *-č*) - the Problem of the Automatic Morphological Analysis

The aim of this paper is to demonstrate how can be used the data mined from corpora for preparation of linguistic basis for NLP (natural language processing) applications. In three representative corpora of literary Czech (SYN2000, SYN2005, SYN2006PUB) the family names (animate masculine on *–č*) were find. The possibility of verbal motivation of them was analyzed thereafter. In this way a list of evantual overgenerations of application of the word formation’s formal rules (Osolsobě 2008) was enlarged.

Seznam literatury

BENEŠ, J. (1962): *O českých příjmeních*. Praha : ČSAV.

BENEŠ, J. (1970): *O českých příjmeních*. Rejstříky. Praha : Zvláštní příloha zpravodaje místopisné komise ČSAV.
BENEŠ, J. (1998): *Německá příjmení u Čechů I., II.* Ústí nad Labem : UJEP.
DOKULIL, M. (1962): *Tvoření slov v češtině.* Praha : Československá akademie věd.
DOKULIL, M., KOMÁREK, M. a kol. (1986): *Mluvnice češtiny I, II.* Praha : Academia.
HAJIČ J. (2004): *Desambiguation of Rich Inflection* (Computational Morphology of Czech). Praha : Karolinum, Charles University Press.
HAVRÁNEK, B. a kol. (1989): *Slovník spisovného jazyka českého (SSJČ)*. Praha : Academia.
HLAVÁČOVÁ, J. (2001): Morphological Guesser of Czech Words. In MATOUŠEK, V. (ed.): *Text, Speech and Dialogue*. Berlin : Springer-Verlag, s. 70-75.
HUJER, O., SMETÁNKA, E., WEINGART, M., HAVRÁNEK, B., ŠMILAUER, V., ZÍSKAL, A. (red.) (1935-1957): *Příruční slovník jazyka českého (PSJČ)*. Praha : Státní nakladatelství.
FILIPEC, J. a kol. (2005): *Slovník spisovné češtiny pro školu a veřejnost (SSČ)*. Praha : Academia.
KARLÍK, P., NEKULA, M., PLESKALOVÁ, J. (2002): Encyklopedický slovník češtiny. Praha : Nakladatelství Lidové noviny.
KLÍMOVÁ, J., ŠTÍCHA, F. (2006): K možnostem a mezím sufixální derivace substantiv. In: ŠTÍCHA, F. (ed): *Možnosti a meze české gramatiky*. Praha : Academia, s. 127-138.
KNAPPOVÁ, M. (2002): Naše a cizí příjmení v současné češtině. Liberec : Tax az Kort.
KOPEČNÝ, F. (1991): *Průvodce našimi jmény*. Praha : Academia.
MATES, V.: (2004): *Jména tajemství zbavená*. Praha : Epocha.
MOLDÁNOVÁ, D. (2004): *Naše příjmení*. Praha : Agentura Pankrác, s.r.o.
OSOLSOBĚ, K. (1996): *Algoritmický popis české morfologie a strojový slovník češtiny.* Brno : FF MU, (disert. práce).
OSOLSOBĚ, K. (2008): Formální popis derivace deverbativ na –č. *SPFFMU A, 56*, s. 121-135.
OSOLSOBĚ K., PALA, K., SEDLÁČEK, R., VEBER, M. (2002): 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,* s. 1254-1259.
PLESKALOVÁ, J. (1998): *Tvoření nejstarších českých osobních jmen*. Brno : Masarykova univerzita.
RYCHLÝ, P. (1998-2003): Bonito – grafické uživatelské rozhraní systému Manatee, Verze 1.49. Dostupné z http://ucnk.ff.cuni.cz/bonito/
SEDLÁČEK, R. (2004): *Morphematic analyser for Czech*. Brno : FI MU, (disert. práce).
SVOBODA, J. (1964): *Staročeská osobní jména a naše příjmení*. Praha : Nakladatelství ČSAV.
ŠMILAUER, V. (1971): *Novočeské tvoření slov*. Praha : Státní pedagogické nakladatelství.
Český národní korpus - SYN2000/SYN2005/SYN2006PUB. Ústav Českého národního korpusu FF UK, Praha 2000. Dostupný z WWW: <http://ucnk.ff.cuni.cz>.
Deriv http://nlp.fi.muni.cz/projekty/ajka/cjbb85/