Klára O s o l s o b ě (Brno) – Hana Ž i ž k o v á (Brno)

Homonymie mezi oikonymy a antroponymy zakončenými na *-slav/-slava* jako problém automatické morfologické analýzy

HOMONYMY AMONG OIKONYMS AND ANTHROPONYMS ENDING IN -SLAV/-SLAVA AS THE PROBLEM OF AUTOMATIC MORPHOLOGICAL ANALYSIS

Homonymy at all levels, which is a distinct feature of all natural languages is also one of the most significant obstacles to automatic natural language processing. In this paper, we will point out the morphosyntactic differences of Czech anthroponyms ending in -slav (Miroslav-type, masculine) and Czech oikonyms with the same ending (Miroslav-type, feminine) and Czech anthroponyms ending in -slava (Miroslava-type feminine, because its forms are homonymous with both: masculine anthroponyms and feminine oikonyms). The analysis of data from the SYN v8 corpus shows that word form homonymy significantly influences the results of automatic morphological analysis. We will document errors in the coverage of the automatic analyzer dictionary and above all errors in morphological tagging and we will propose a solution to partially improve the automatic disambiguation of the given type of proper nouns.

Klíčová slova: antroponymum, oikonymum, zakončení na -slav/-slava, korpus, morfologické značkování, desambiguace,

Key words: anthroponym, oikonym, ending in -slav/-slava, corpus, tagging, disambiguation

Úvod

Homonymie na všech úrovních, která je nepřehlédnutelnou vlastností každého přirozeného jazyka, je zároveň jednou z nejvýraznějších překážek automatického zpracování přirozeného jazyka. V tomto příspěvku poukážeme na morfosyntaktické odlišnosti antroponym zakončených na *-slav* (maskulina typu *Miroslav*), oikonym se stejným zakončením (feminina typu *Miroslav*) a antroponym zakončených na *-slava* (feminina typu *Miroslava*, jejichž tvary jsou homonymní s antroponymy maskuliny i s oikonymy femininy). Na analýze dat z korpusu SYN v8 ukážeme, že tvarová homonymie značně ovlivňuje výsledky automatické morfologické analýzy. Doložíme chyby v pokrytí slovníku automatického analyzátoru a především chyby v morfologickém značkování a navrhneme řešení na částečné zlepšení automatické desambiguace uvedeného typu vlastních jmen.

Historický exkurz

Významným pojmenovacím motivem, na jehož základě se tvořila místní jména od samého počátku češtiny, byla posesivita. Vlastní jméno posesivní vyjadřuje vlastnictví objektu konkrétní osobou nebo nositelem určité funkce. Je utvořeno z antroponyma, popř. apelativa, a to nejčastěji sufixy *-ov-* nebo *-ín-*. Skupina vlastních jmen posesivních zahrnuje typologicky pestrou škálu oikonym tvořených z antroponym, popř. apelativních označení osob

reprezentujících určité funkce, např. *Vítkov*, *Opatov*, tj. ‘Vítkův/opatův majetek, hrad n. dvůr’. (David, 2017).

Vedle převažujícího sufixu *-ov* se ale též uplatňoval praslovanský sufix **-jь*. V praslovanštině se spojoval jen s názvy živých bytostí (stsl. *člověčь*), jisté omezení bylo možné pozorovat u základů jo-kmenů, jejichž poslední konsonant nepodléhal alternaci (*-lь* apod.) Tato derivace se totiž už od praslovanských dob (od splynutí -j- s posledním konsonantem základu) manifestuje jen konsonantickou alternací, a proto by neměla smysl, srov. *roditelь* (subst.) > **roditel´ь* (adjektivum, ale rovněž jmenná deklinace). Z tohoto důvodu sufix **-jь* v rané staré češtině svou produktivitu ztrácí (Lamprecht – Šlosar – Bauer, 1986). Po určitou dobu se však sufix **-jь* uplatňuje při tvoření posesivních místních jmen jako například např. *Boleslav*, *Břeclav*, *Chotěboř*, *Jaroměř*, *Litomyšl*, *Mladeč*, *Spytihněv*, tj. ‘Boleslavův’ (majetek, hrad n. dvůr)(David, 2017).

Co se týče přechodu od původně mužského rodu k rodu ženskému, Šrámek a Lutterer (2004) uvádějí, že právě tato starobylá přivlastňovací přípona *-jb* způsobila změkčení koncového konsonantu a podmínila přechod od původně rodu mužského (ten Břěcislav, hrad Břetislavův) k rodu ženskému (ta Břeclav – srov. Stará a Mladá Boleslav). David s Rousem (2006) se zase domnívají, že přechod jmen od mužského rodu k ženskému byl způsoben patrně vlivem velkých skupin místních jmen ženského rodu jako jsou jména typu *Lomnice, Roudná a Litoměřice*.

V dnešní češtině jsou naprosto bez výjimek antroponyma zakončená na -slav rodu mužského a místní jména zakončená na -slav rodu ženského.

Substantiva zakončená na *-slav/-slava* v synchronních korpusech a ve slovníku *MorfFlex*

V následující analýze se opíráme o data získaná z korpusu syn v8 (spojení korpusů řady SYN, verze 8 z 12. 12. 2019, velikost: 5 391 362 082 pozic). (Křen a kol., 2019) Data jsou morfologicky značkována, použitý byl slovník *Morfflex* (Hajič – Hlaváčová, 2016) a další automatické nástroje[1].

Substantiva *Bohuslav, Boleslav, Borislav, Bořislav, Břetislav, Budislav, Domaslav, Jaroslav, Miroslav, Přibyslav, Soběslav, Sulislav, Svatoslav, Vladislav, Vlastislav, Vratislav, Zbraslav, Zbyslav, Zdeslav, Zdislav* mají ve slovníku *MorfFlex*[2] jak interpretaci antroponym (maskulin životných), tak oikonym (feminin) a většina má protějšky mezi femininy antroponymy zakončenými na *-slava*.

Ve slovníku *MorfFlex* jsou pouze jako feminina oikonyma označena substantiva *Budislav G*[3]*, Čáslav G*[4]*, Nadslav G, Nosislav G, Perejaslav G, Předslav G, Preslav G,*

[1] Viz https://wiki.korpus.cz/doku.php/pojmy:morfologicka_analyza.

[2] Přestože morfologický slovník *MorfFlex* (Hajič – Hlaváčová, 2016) zahrnuje kromě morfologické značky (tagu) také informace o vlastních jménech a rozlišuje křestní jména (_Y), příjmení (_S) a geografické názvy (_G), není podle autorů slovníku v těchto interpretacích ani důsledný, ani zcela spolehlivý. Z tohoto důvodu jsme se pokusili formulovat co nejvíce pravidel pro desambiguaci, aniž bychom se o příslušné informace opírali.

[3] *Příjmení Jozefa Zieleňce (či Zieleniece?) má kořeny v polštině, Petry Buzkové vzniklo z osobního jména* ***Budislav/Budislav/NNFS1-----A-----****.*

[4] *… Jméno vsi je odvozeno od středověkého jména Půta; již r. 1183 je uveden v listinách jakýsi* ***Čáslav/Čáslav/NNFS1-----A-----*** *z Potína;*

Sudislav G[5]*, Sulislav G*[6]*, Vrchoslav G, Wodzislav G, Zbraslav G*[7], nicméně v korpusech se okrajově některá vyskytují i jako antroponyma (viz doklady v poznámkách vč. interpretací přiřazených automatickou morfologickou analýzou, které nejsou s ohledem na nedostatečné pokrytí slovníku výsledkem chybné desambiguace[8]).

Ve slovníku *MorfFlex* figurují pouze jako antroponyma maskulina jména *Blahoslav Y, Boguslav Y, Bojislav Y, Božislav Y, Břetislav Y, Bronislav Y, Čechoslav Y, Česlav Y, Dobroslav Y*[9]*, Drahoslav Y, Hostislav, Y, Kostislav Y, Krunoslav Y, Květoslav Y, Ladislav Y, Levoslav Y, Liboslav Y, Luboslav Y, Miloslav Y, Mnislav Y, Pobraslav Y, Pravoslav Y, Radoslav Y, Radslav Y, Rastislav Y, Rostislav Y, Stanislav Y*[10]*, Stranislav Y, Svjatoslav Y Vácslav Y, Věnceslav Y, Věroslav Y, Věslav Y, Věstislav Y, Vítězslav Y, Vitoslav Y, Vjačeslav Y, Vjekoslav Y, Vrastislav Y, Vseslav Y, Zbislav Y*[11]*, Zděslav Y, Zlatoslav Y*[12]*,* nicméně v korpusech se okrajově některá vyskytují i jako oikonyma (viz doklady v poznámkách vč. interpretací přiřazených automatickou morfologickou analýzou, které nejsou s ohledem na nedostatečné pokrytí slovníku výsledkem chybné desambiguace).

Jako antroponyma feminina jsou ve slovníku *MorfFlex* interpretována substantiva *Miroslava, Stanislava, Miloslava, Květoslava, Ladislava, Zdislava, Čáslava*[13]*, Vladislava, Bohuslava, Vítězslava, Bronislava, Drahoslava, Věnceslava, Boleslava, Dobroslava, Vlastislava, Svatoslava, Přibyslava, Střezislava*[14]*, Liboslava, Bořislava, Hostislava*. Ke všem ovšem existují i antroponyma maskulina životná, k řadě z nich i feminina oikonyma na *slav*. Substantiva *Bratislava, Oslava, Úslava, Sudslava,* jsou ve slovníku *MorfFlex* interpretována pouze jako feminina oikonyma/hydronyma. V korpusech se ovšem vyskytují i antroponyma maskulina *Bratislav*[15] a *Sudslav*[16].

Slovník *MorfFlex* vykazuje nedostatky v pokrytí proprií zakončených na *-slav/-slava.* Navrhujeme, aby do slovníku *MorfFlex* byla doplněna: a) antroponyma maskulina životná

---

[5] *První zmínka o zdejším sídle je z roku 1285, kdy zde měl majetek* ***Sudislav/Sudislav/NNFS1-----A-----*** *ze Zásmuk, královský kuchmistr.*

[6] *První zmínka o obci Zásmuky se datuje k roku 1285, kdy byl majitelem zdejšího panství* ***Sulislav/Sulislav/NNFS1-----A-----*** *ze Zásmuk.*

[7] *Po sto letech se však jako držitel vsi připomíná okolo roku 1238* ***Zbraslav/Zbraslav/NNFS1-----A-----*** *z Miletína, významný šlechtic na dvoře krále Václava I.*

[8] Viz https://wiki.korpus.cz/doku.php/pojmy:desambiguace.

[9] *… Bájo představení Cirkus* ***Dobroslav /Dobroslava/NNFP2-----A-----*** *…* . Další kolokace označené jako NNFP2.* (femininum v genitivu plurálu) jsou např. *… v časopisu* ***Dobroslav****..., … akademický spolek* ***Dobroslav*** *..., …* Domníváme se, že v těchto případech jde spíše o maskulina neživotná. Je ovšem otázka, zda je třeba vícerou interpretací ve slovníku zatěžovat desambiguaci.

[10] *… Ve stejné době se u města* ***Stanislav/Stanislav/NNMS1-----A-----*** *nacházela německá 1 . a 8 . tanková divize ...*

[11] *… Kapli v osadě* ***Zbislav/Zbislav/NNMS1-----A-----*** *opraví obec Zhoř ...*

[12] *… řekla vedoucí laborantka* ***Zlatoslava /Zlatoslav/NNMS2-----A-----*** *Přecechtělová.*

[13] V korpusech SYN v8 je ovšem doloženo hojně příjmení *Čáslava*. Možnost femininní interpretace mají kolokace *společnost/firma/hospoda/penzion/polosamota/chata Čáslava* (přestože je tvar tagovaný jako femininum, může jít stejně dobře o maskulinum, podle kontextu jde o nominativ jmenovací) a z dokladů na internetu vyplývá, že jde o příjmení majitele firmy. Na relevantní doklady křestního jména *Čáslava* jsme v korpusu SYN v8 nenarazili.

[14] Slovník *MorfFlex* ovšem má jen antroponymum femininum *Střezislava*, navzdory v korpusu SYN v8 okrajově doloženému maskulinu *Střezislav.*: *… Na zámek přijel* ***Střezislav /Střezislava/NNFP2-----A-----*** *Megawat , proslavený šermíř ...*

[15] *… přiletěli si pro něj srbský ministr obrany* ***Bratislav/Bratislava/NNFP2-----A-----*** *Gašič …, … říká v rozhovoru pro MF DNES archeolog* ***Bratislav/Bratislava/NNFP2-----A-----*** *Resutík ...*

[16] *… Poprvé je o ní zmínka v Zemských deskách z r. 1263, kdy o ní psal rytíř* ***Sudslav/Sudslava/NNFP2-----A-----*** *z Dubného ...*

(***Budislav, Bratislav, Čáslav, Sudislav, Sudslav, Sulislav, Střezislav, Zbraslav***), b) oikonyma feminina (***Stanislav, Zbislav***), c) antroponyma feminina (***Zlatoslava***) a přimlouváme se za d) vyřazení feminina antroponyma ***Čáslava***. Návrh se opírá o doklady v textech v korpusu SYN v8.

Nedostatečné pokrytí slovníku se ovšem týká i dalších proprií zakončených na *-slav/-slava*. Lze je získat z dat korpusu SYN v8 dotazem [word="[[:upper:]].*(slav|slava)" & tag="X.*"]. Množství z takto vyhledaných dat jsou překlepy, spojená slova atd. Objevují se ovšem i relevantní propria. S frekvencí 5 a více výskytů jsou v korpusu SYN v8 doložena antroponyma – křestní jména v mužské i ženské podobě, většinou pocházejí z různých slovanských jazyků, často jde o grafické varianty.[17] Kromě skutečných antroponym jsou doložena žertovná jména (*Gumislav, Pepislav, Bečislav, Šlaposlav, Knihoslav, Dluhoslav*[18]*, Pravdoslav, Hromoslav*), příjmení (*Latislav, Ratislav, Záslav, Pěvoslav, Vrchoslav*), oikonyma (*Sukoslav, Stejslav*[19]*, Donuslav, Tvrdoslav, Trojslava, Předslava*), hydronymum *Reslava*. O produktivitě modelu svědčí hravé názvy jako jméno souboru *Pěslav,* kapela *Smrtislav,* divadelní spolek *Milislav,* pivo *Kvasslav,* sportovní tým ragbistek *Pepřeslav* (jména by patrně měla tvary neživotných maskulin).

Všechna lemmata navrhujeme zařadit do slovníku *MorfFlex.*

Homonymní tvary proprií zakončených na *-slav/-slava*

Pro automatickou morfologickou analýzu není ovšem důležité jen pokrytí slovníku. Kvalitní anotace je neméně závislá na přesnosti desambiguace. Homonymní tvary antroponym zakončených na *-slav/-slava* a oikonym zakončených na *-slav* představují velký problém pro spolehlivou desambiguaci.

Jestliže slovní tvar začíná velkým písmenem a končí na *-slav* (formulováno v jazyce CQL word="[[:upper:]].*slav"), pak mohou nastat tyto případy lemmatizace a morfologické interpretace vyjádřené v podobě značek užívaných při značkování korpusů ČNK:
a) lemma=".*slav" & tag="NNFS[14].*"
b) lemma=".*slav" & tag="NNMS1.*"
c) lemma=".*slava" & tag="NNFP2.*"

Jestliže slovní tvar začíná velkým písmenem a končí na *-slava* (word="[[:upper:]].*slava"), pak mohou nastat tyto případy:
d) lemma=".*slav" & tag="NNMS[24].*"
e) lemma=".*slava" & tag="NNFS1.*"

---

[17] Seznam: *Vladyslav(a), Víťazoslav, Ludoslav(a), Czeslav, Vojslav(a), Hviezdoslav, Hvězdoslav(a), Mojslav(a), Krasoslav(a), Ljuboslav, Svetoslav, Mislav, Milislav, Ľuboslav(a), Myroslav(a), Kvetoslav(a), Ventzislav, Václav, Světoslav(a), Přemyslav, Vyacheslav, Yaroslav, Věkoslav(a), Wladyslav(a), Šlechtislav, Viacheslav, Svetislav, Wieslav(a), Vitislav, Lubislav(a), Ladoslav, Vencislav, Ninoslav, Rostlislav, Vieroslav(a), Berislav, Držislav, Veaceslav, Uladzislav, Wenceslav , Wladislav(a), Venceslav, Vítoslav, Przemyslav, Mnohoslav, Zorislav, Věřislav, Mečislav, Rostyslav, Něhoslav(a), Žiarislav, Světislav, Leslav, Svéslav, Vněslav, Kristoslav, Bogislav(a), Čistoslav, Vlaslav, Chotěslav, Novislav, Siloslav, Ruslav, Ludislav(a), Zveneslav, Prvoslav, Sviatoslav, Byslav, Pivoslav, Lavoslav, Beloslav, Lutoslav, Unislav, Tsvetoslav, Svatislav, Doběslav, Veleslav(a), Mieczyslav, Vidoslav, Předislav, Zlatislava, Desislava, Kvitoslava, Cvetislava, Wislava, Preslava*.
[18] Substantivum je hravou narážkou na ekonomickou politiku *Bohuslava Sobotky*.
[19] V korpusu SYN v8 je v kontextu *… na hřišti* ***Stejslav*** *Hýskov …* Na internetu je doloženo i antroponymum: *… 1185 vlastnil* ***Stejslav*** *z Nitrianskej Blatnice …*(Koplotovce, 2020).

Jestliže slovní tvar začíná velkým písmenem a končí na *-slavy* (word="[[:upper:]].*slavy"), pak mohou nastat tyto případy:

f) lemma=".*slav" & tag="NNMP[47].*"

g) lemma=".*slava" & tag="NNF(S2|P[14].*"

Konkrétní případ homonymie uvnitř skupiny proprií zakončených na *-slav/-slava* je přehledně představen v následující tabulce 1.

Tabulka 1

lemma=	"**[[:upper:]]**.***slav**" &		"**[[:upper:]]**.***slav**" &	"**[[:upper:]]**.***slav**" &	"**[[:upper:]]**.***slava**" &	"**[[:upper:]]**.***slava**" &	"**[[:upper:]]**.***slava**" &
tag=	"NN**M**S1.*"		"NN**F**S1.*"	"NN**F**S4.*"	"NN**F**P2.*"		
word="[[:upper:]].*slav"	Jak informoval **Miroslav** Chlan ...		… patří do katastru města **Miroslav** ...	V opačném směru bude provoz veden přes **Miroslav** do Hostěradic.	To už musí posoudit manželé a přátelé všech **Miroslav**.[20]		
tag=	"NN**M**S2.*"	"NN**M**S4.*"			"NN**F**S1.*"		
word="[[:upper:]].*slava"	Černobílá kolekce fotografií **Miroslava** Kolbasy.	Jmenovitě jde o **Miroslava** Tarnóczyho.			… vedoucí muzea **Miroslava** Štýbrová.		
tag=	"NN**M**P4.*"	"NN**M**P7.*"			"NN**F**S2.*"	"NN**F**P1.*"	l"NN**F**P4.*"
word="[[:upper:]].*slavy"	… budou muset schovat své **Miroslavy**, Kalouska a Šloufa ...	… hraje spolu s **Miroslavy** Šimandly - otcem a synem ...				… bylo nás tam pět rodiček a z toho dvě **Miroslavy**[21] úplně stejného věku ...	… Kdybychom měli vyjmenovat všechny **Miroslavy**[22], které se nějak proslavily, …

Zajímavý doklad tvarové homonymie nacházíme v následujícím textu: *Den s Mírou ve Veltrusech sobota 10 . března Veltruští zámečtí zvou všechny **Miroslavy**, Míry i Mirky, aby oslavili svůj svátek poněkud netradičně. Pokud jste **nositelem/nositelkou** jména **Miroslav / Miroslava**, přijeďte do Veltrus a vydejte se na komentovanou prohlídku s názvem Váhy, míry, závaží na zámku vám ukáží …* Tvar ***Miroslavy*** je co do gramatického rodu nedesambiguovatelný, neboť označuje osoby obou pohlaví. V desambiguační praxi užívané při anotování korpusů se v analogických případech dává přednost hierarchicky neutrálnější interpretaci (maskulinum životné).

S ohledem na rozsah článku se budeme nadále zabývat pouze problematikou homonymních tvarů zakončených na *-slav/-slava*, tedy případy a) až e). Dokladů plurálových

[20] Uvedený doklad nepochází z korpusu SYN v8, je upravený pro názornou představu. Dokladů v korpusu SYN v8 je poměrně málo, ale vyskytují se: *… manželé a přátelé všech **Drahoslav**, Drahun a Drahušek …, … Patronkou všech **Zdislav**, ...*

[21] Uvedený doklad nepochází z korpusu SYN v8, je upravený pro názornou představu. Ve skutečnosti se v korpusu SYN v8 vyskytuje takto: *… bylo nás tam pět rodiček a z toho dvě **Jaroslavy** úplně stejného věku ...*

[22] Doklad jsme jednoduše v celém korpusu SYN v8 nedokázali najít, což nevylučuje jeho výskyt. V korpusu Araneum jsme našli jeden relevantní doklad (*… Kdybychom měli vyjmenovat všechny Jaroslavy, které se nějak proslavily, trvalo by to do Silvestra …*), který jsme pro názornost upravili.

tvarů proprií není mnoho, přesto se v korpusech vyskytují a nejsou vždy správně desambiguované. Pokládáme tudíž za užitečné se o nich alespoň zmínit.

Zběžnou rešerší v korpusu SYN v8 (dotazem na slovní tvar začínající velkým písmenem a končící na *-slav* nebo *-slava* s vyloučením apelativního substantiva *oslava*[23] [word="[[:upper:]].*(slav|slava)" & lemma!="oslava"]) získáme celkem 3 716 726 výskytů tvarů sledovaných proprií. V desambiguaci není na první pohled mnoho chyb, nicméně zapátráme-li u jednotlivých slov, chyby se začnou objevovat[24]. Z tohoto důvodu jsme se zaměřili na případy, které by bylo možné správně desambiguovat pomocí pravidel. Pravidla lze opřít o grafické, morfosyntaktické a lexikální vlastnosti kontextu, v němž se slovní tvary končící na *-slav/-slava* vyskytují v korpusech.

Kolokace a víceslovná pojmenování, která lze zařadit do databáze *Lemur*[25]

Poměrně dobře by bylo možné vylepšit desambiguaci v případech, kdy proprium (antroponymum) zakončené na *.*slav* je v kontextu příjmení významné osoby, respektive případy, kdy se kombinace slovo zakončené na *-slav/-slava* vyskytuje bezprostředně před slovním tvarem s velkým počátečním písmenem, které lze interpretovat jako příjmení. Typickým případem je *Bohuslav Martinů*. V korpusu SYN v8 je celkem 2 200 výskytů spojení ***Bohuslav Martinů***. V 1903 případech je tvar *Bohuslav* interpretován jako životné maskulinum v nominativu singuláru. 297 má interpretaci jako femininum (jak oikonymum s lemmatem *Bohuslav,* tak antroponymum *Bohuslava*). V témže korpusu je celkem 16 976 výskytů spojení ***Bohuslava Martinů.*** V 608 případech je tvar *Bohuslava* interpretován jako nominativ singuláru feminina *Bohuslava.*

V případě spojení ***Jaroslav Seifert*** jsou mylně jako tvary oikonyma interpretovány pouze 2 doklady. V případě spojení ***Jaroslava Seiferta*** je mylně jako tvar feminina *Jaroslava* interpretován 1 doklad. V případě spojení ***Miroslav Horníček*** jsou všechny tvary interpretovány správně. V případě spojení ***Miroslava Horníčka*** je mylně jako tvar feminina *Miroslava* interpretován 1 doklad.

Dotazem [word="[[:upper:]].*slav"][word="[[:upper:]].*"] lze získat další kandidáty. ***Bohuslav Sobotka/Bohuslava Sobotky*** (90 feminin), ***Miroslav Kalousek/Miroslava Kalouska*** (1 femininum), ***Miroslav Pelta/Miroslava Pelty*** (5 feminin), ***Miroslav Donutil/Miroslava Donutila*** (5 feminin), ***Stanislav Gross/Stanislava Grosse*** (1 femininum), ***Jaroslav Šilhavý/Jaroslava Šilhavého*** (4 feminina). S ohledem na poměrně dobré výsledky desambiguace u frekventovaných takto vyhledaných jmen jsme se pokusili zformulovat obecnější pravidlo pro feminina.

---

[23] Apelativum *oslava* je homonymní s hydronymem *Oslava*. Tuto homonymii automatická morfologická analýza zahrnuje, a sice prostřednictvím dvou lemmat (s velkým a malým počátečním písmenem), tvarová homonymie není problém, neboť tvarosloví apelativa a propria se neliší. Problémem nepříliš spolehlivé desambiguace lemmatu *oslava/Oslava* se na tomto místě nebudeme zabývat. Přesahuje hranice i možnosti automatické morfologické analýzy.

[24] Například vyfiltrujeme-li případy, kdy je tvar *Jaroslav* interpretován jako femininum – oikonymum, narazíme dosti často na kolokace typu *… manželé Dana a **Jaroslav /Jaroslav/NNFS1-----A-----** Stodolovi …*, které jsou se zvolenou interpretací naprosto v rozporu, podobně je tomu i v kolokacích propria *Bohuslav Martinů, Bohuslav Sobotka* atd.

[25] Databáze víceslovných lexikálních jednotek LEMUR (Multiword Expression Lexical Database) vzniká péčí Ústavu teoretické a komputační lingvistiky FF UK a Ústavu Českého národního korpusu a využívá se při desambiguaci korpusů Českého národního korpusu (např. Jelínek a kol., 2018).

Testovali jsme dotaz na tvar *.*slava* v kontextu proprií příjmení na *.*ová.* Jednoduchým dotazem [word="[[:upper:]].*slava" & tag!="..FS1.*"] [word="[[:upper:]].*ová"] lze získat 13086 dokladů, z toho 13085 chybně desambiguovaných proprií zakončených na *.*slava*. Jediný správně označkovaný doklad je příjmení *Čáslava* (*... Snímky : Karel **Čá<u>slava</u> N<u>ová</u** dětská hřiště jsou již v provozu*). Vzhledem k tomu, že k příjmení *Čáslava* by nemělo být ve slovníku uvedeno jako femininum antroponymum *Čáslava*, ale pouze jako maskulinum *Čáslava* a femininum oikonymum *Čáslav*, stačilo by při desambiguaci zohlednit jednoduché kontextové pravidlo a odstranit tak další tisícovku chyb.

Podobně koncipovaný dotaz pro maskulina ([word="[[:upper:]].*slav" & tag!="..MS1.*"][word="[[:upper:]].*[^á]"]) dával poměrně přegenerované výsledky. Nicméně frekvenční distribuce slovních tvarů nejvíce vpravo pomohla najít další adepty na spojení rodné jméno + příjmení, které lze uložit do databáze ***Lemur***[26] (***(Miro|Jaro|Bohu|Vladi)slav Hořejší*** 313 chyb, ***Vladislav Lipovský*** 258 chyb, ***Jaroslav Krček*** 210 chyb, ***Jaroslav Mokrý*** 186 chyb, ***Jaroslav Hlavnička*** 128 chyb, ***Jaroslav Doubrava*** 123 chyb a mnoho dalších dvojic s frekvencí nižší než 100).

Podobně lze postupovat i u oikonym jako ***Mladá Boleslav***, ***Stará Boleslav***. Dotazem [word="(Mlad(á|ou))|(Star(á|ou))"][word="Boleslav" & tag!="..FS[14].*"] lze získat 46 chybně desambiguovaných tvarů.[27]

Návrh pravidel desambiguace proprií zakončených na *-slav/-slava* v koordinovaných skupinách

Všimli jsme si, že se antroponyma zakončená *na -slav/-slava* velice často vyskytují v kontextu *manžel[ée]/rodiče antropnymum [MF] a antropnymum [MF] proprium na ov[iy].* Konstatovali jsme, že ne vždy je desambiguace antroponym správně. Velmi často je tvar maskulina antroponyma zakončeného na *-slav* interpretován jako tvar feminina oikonyma zakončeného na *-slav*, méně často pak je tvar antroponyma feminina zakončeného na *-slava* interpretován jako tvar antroponyma maskulina zakončeného na *-slav* a naopak.

Navrhli jsme kontextová pravidla pro desambiguaci. Některá pravidla počítají s pravopisně správnými tvary v okolí analyzovaného tvaru. Korpusy ovšem zahrnují i texty s pravopisnými chybami. Těmi se v následující analýze systematicky nezabýváme, pouze na ně okrajově upozorňujeme.

Poznamenáváme, že pravidla týkající se dvojice s antroponymy zakončenými na *-slav/-slava* se týkají i dvojic s antroponymy *Václav/Václava* (ale i dalších podobných, jako například *.*mil/.*mila, .*mír/.*míra*, tedy *Bohumil/Bohumila, Jaromír/Jaromíra* atd.). Vzhledem k rozsahu tohoto článku je ponecháme stranou.

Pravidla zahrnující v podmínce tvary *rodiče/manžel[ée]* je třeba nazírat s následující výhradou. Tvar ***rodiče*** má homonymní tvar pro nominativ a akuzativ, navíc může být

---

[26] Většina uvedených příjmení mohou být i apelativa (adjektivum *hořejší,* apelativní substantiva *krček, doubrava, hlavnička*). Z toho důvodu se nelze spoléhat na značkování proprií ve slovníku *MorfFlex,* neboť i když jsou příjmení uložena jako propria, jejich značkování je předmětem desambiguace.

[27] Testovali jsme rovněž dotaz na další víceslovná oikonyma [word="(Velk(á|ou))|(Mal(á|ou)|Horní|Dolní)"][word=".*slav" & tag!="..FS[14].*"], ale relevantní doklady jsme nenašli. Rovněž jsme testovali dotaz na případy potenciálních víceslovných oikonym s prvním slovem *Mladá* a *Stará*, s vyloučením druhého slova *Boleslav* (dotaz: [word="(Mlad(á|ou))|(Star(á|ou))"][word=".*slav" & word!="Boleslav" & tag!="..FS[14].*"]). Kromě 83 překlepů ve slově *Boleslav* jsme našli pouze 4 nerelevantní doklady (*... ředitel Domova **Mladá Miloslav /Miloslav/NNMS1-----A-----** Müller ...*).

rozvinut postponovaným (i koordinovaným) přívlastkem v genitivu. Máme-li tvar rodiče následovaný dvojicí potenciálních proprií, mezi nimiž je slučovací spojka *a*, pak můžeme předpokládat tyto interpretace: *rodiče **Stanislava** a **Miroslava*** ← (*Stanislavov[iy]) a Miroslavov[iy]) rodiče*) | ((*Stanislavovi rodiče a Miroslava/Fem.*) | (*Stanislavovy rodiče a Miroslava/Mask.*) | (*rodiče/Nom. Stanislava/Fem. a Miroslava/Fem.*) | (*rodiče/ak. Stanislava/Mask. a Miroslava/Mask.* Poslední dvě interpretace jsou interpretace počítající s rodičovským stejnopohlavním párem). Na tyto případy upozorňujeme, přestože některé z interpretací jsou velmi málo pravděpodobné a pravidla, která navrhujeme, se takovými případy nezabývají.

Pravidla jsou formulována tak, aby se neopírala o automatickou morfologickou analýzu tvarů v kontextu slova na *slav/slava*. V pravidlech počítáme s variantami kombinací antroponym žen a mužů, přičemž vždy jeden člen dvojice má explicitně tvar zakončený na *slav/slava*. V pravidlech 1, 1a předpokládáme, že před *a* stojí femininum, nicméně se může vyskytovat i maskulinum a za *a* maskulinum na *slav*. V pravidlech 2, 2a předpokládáme, že před *a* stojí maskulinum na *slav* a po *a* stojí femininum, nicméně se může vyskytovat i maskulinum. V pravidlech 3, 3a, předpokládáme, že před *a* stojí femininum na *slava* a za *a* maskulinum, kdežto v pravidlech 4, 4a je pořadí opačné. V pravidlech 5, 5a předpokládáme, že před *a* stojí maskulinum (tvar na *slava*) v akuzativu a za *a* femininum, kdežto v pravidlech 6, 6a je pořadí opačné. Je patrné, že preferované slovosledné varianty (v koordinované skupině je na prvním místě jméno ženy a na druhém jméno muže - odraz zdvořilostní strategie) jsou početnější, zachytíme v nich tudíž většinou i více chyb v desambiguaci.

1. Jestliže KWIC je **[word="[[:upper:]].*slav"& tag="NN(MS1|FS1|FS4|FP2).*"]**
pak v případě, že levý kontext <-2,-1> je [word="[[:upper:]].*"][lc="a"]
KWIC je **[word="[[:upper:]].*slav"]**
a pravý kontext <1,1> je [word="[[:upper:]].*ovi"]
pak platí, že **[word="[[:upper:]].*slav"& tag="NNMS1.*" & lemma="[[:upper:]].*slav"]**
Tomuto pravidlu odporují výsledky desambiguace tvaru propria zakončeného na *-slav* v korpusu SYN v8 v 1417 případech, viz pravidlo 1. v tabulce 2 níže.

1a. Jestliže **[word="[[:upper:]].*slav"& tag="NN(MS1|FS1|FS4|FP2).*"]**
pak v případě, že levý kontext <-3,-1> je [lc=".*rodiče|.*manželé"][word="[[:upper:]].*"][lc="a"]
KWIC je **[word="[[:upper:]].*slav"]**
a pravý kontext <1,1> je [word!="[[:upper:]].*ovi"]
pak velmi pravděpodobně[28] platí, že **[word="[[:upper:]].*slav"& tag="NNMS1.*" & lemma="[[:upper:]].*slav"]**
Tomuto pravidlu odporují výsledky desambiguace tvaru propria zakončeného na *-slav* v korpusu SYN v8 v 212 případech, viz pravidlo 1a. v tabulce 2 níže.

2. Jestliže KWIC je **[word="[[:upper:]].*slav"& tag="NN(MS1|FS1|FS4|FP2).*"]**
pak v případě, že pravý kontext <1,3> je[lc="a"] [word="[[:upper:]].*"] [word="[[:upper:]].*ovi"]
pak platí, že **[word="[[:upper:]].*slav"& tag="NNMS1.*" & lemma="[[:upper:]].*slav"]**

[28] S velmi málo pravděpodobnou interpretací, jako např. *rodiče Martina a Jaroslav ← Martinovi rodiče, kteří jsou zároveň rodiči několika Jaroslav*, nepočítáme. Bereme v úvahu pouze pravděpodobnější interpretace (*rodiče Martina a Jaroslav ← rodičovský pár Martina/ žena a Jaroslav / muž | Martinovi rodiče a Jaroslav/ muž*).

Tomuto pravidlu odporují výsledky desambiguace tvaru propria zakončeného na *-slav* v korpusu SYN v8 v 137 případech, viz pravidlo 2. v tabulce 2 níže.

2a. Jestliže KWIC je **[word="[[:upper:]].*slav"& tag="NN(MS1|FS1|FS4|FP2).*"]**
pak v případě, že levý kontext <-1,-1> je [lc=".*rodiče|.*manželé"]
KWIC je **[word="[[:upper:]].*slav"]**
a pravý kontext <1,3> je [lc="a"] [word="[[:upper:]].*"] [word!="[[:upper:]].*ovi"]
pak velmi pravděpodobně[29] platí, že **[word="[[:upper:]].*slav"& tag="NNMS1.*" & lemma="[[:upper:]].*slav"]**
Tomuto pravidlu odporují výsledky desambiguace tvaru propria zakončeného na *-slav* v korpusu SYN v8 v 1 případě, viz pravidlo 2a. v tabulce 2 níže.

3. Jestliže KWIC je **[word="[[:upper:]].*slava"& tag="NN(MS2|MS4|FS1).*"]**
pak v případě, že pravý kontext <1,3> je [lc="a"] [word="[[:upper:]].*[^aeěiouyáéóůý]"] [word="[[:upper:]].*ovi"]
pak velmi pravděpodobně platí, že **[word="[[:upper:]].*slava"& tag="NNFS1.*" & lemma="[[:upper:]].*slava"]**
Tomuto pravidlu odporují výsledky desambiguace tvaru propria zakončeného na *-slava* v korpusu SYN v8 v 139 případech, viz pravidlo 4. v tabulce 2 níže.

3a. Jestliže KWIC je **[word="[[:upper:]].*slava"& tag="NN(MS2|MS4|FS1).*"]**
pak v případě, že levý kontext <-1,-1> je [lc=".*rodiče|.*manželé"]
KWIC je **[word="[[:upper:]].*slava"]**
a pravý kontext <1,3> je [lc="a"] [word="[[:upper:]].*[^aeěiouyáéóůý]"][word!="[[:upper:]].*ovi"]
pak velmi platí, že **[word="[[:upper:]].*slava"& tag="NNFS1.*" & lemma="[[:upper:]].*slava"]**
Tomuto pravidlu odporují výsledky desambiguace tvaru propria zakončeného na *-slava* v korpusu SYN v8 v 65 případech, viz pravidlo 4a. v tabulce 2 níže.[30]

4. Jestliže KWIC je **[word="[[:upper:]].*slava"& tag="NN(MS2|MS4|FS1).*"]**
pak v případě, že levý kontext <-2,-1> je [word="[[:upper:]].*[^aeěiouyáéóůý]"][lc="a"]
KWIC je **[word="[[:upper:]].*slava"]**
a pravý kontext <1,1> je [word="[[:upper:]].*ovi"]
pak velmi pravděpodobně platí, že **[word="[[:upper:]].*slava"& tag="NNFS1.*" & lemma="[[:upper:]].*slava"]**
Tomuto pravidlu odporují výsledky desambiguace tvaru propria zakončeného na *-slava* v korpusu SYN v8 v 21 případech, viz pravidlo 3. v tabulce 2 níže.

---

[29] S velmi málo pravděpodobnými interpretacemi, jako např. *rodiče Jaroslav a Martina* ← (*rodiče několika Jaroslav, kteří jsou zároveň Martinovými rodiči* | (*rodiče/ak. několika Jaroslav a Martina/Mask. ak.* | *rodiče/nom. několika Jaroslav a Martina/Fem. nom.*) nepočítáme.

[30] Pravidla 3., 3a., 4., 4a. jsou formulována tak, že zachycují pouze chyby v kontextu maskulin, která v nominativu singuláru končí na souhlásku, nebo na *í* (tedy nikoli jména jako *Ivo, René, Arne, Ota, Harry, Joe, Otto, Henry, José,* …, ale ani hypokoristika jako *Jirka, Kája, Standa, Pepa, Franta, Ondra,* …, přičemž mnohá z nich jsou homonymní s feminini - *Vlasta, Stáňa, ...*). Například v dokladu … *Radují se z něj* ***manželé /manžel/NNMP1-----A----- Zdislava /Zdislav/NNMS2-----A----- a /a/J^---------- Yassine /Yassin/NNMS5-----A-----*** *z Ústí nad Orlicí* … nebude chyba desambiguace tvaru *Zdislava* pravidlem zachycena (true negative).

4a. Jestliže KWIC je **[word="[[:upper:]].*slava"& tag="NN(MS2|MS4|FS1).*"]**
pak v případě, že levý kontext <-3,-1> je
[lc=".*rodiče|.*manželé"][word="[[:upper:]].*[^aeěiouyáéóůý]"][lc="a"]
KWIC je **[word="[[:upper:]].*slava"]**
a pravý kontext <1,1> je [word!="[[:upper:]].*ovi"]
pak velmi pravděpodobně platí, že **[word="[[:upper:]].*slava"& tag="NNFS1.*" & lemma="[[:upper:]].*slava"]**
Tomuto pravidlu neodporují žádné výsledky desambiguace tvaru propria zakončeného na *-slava* v korpusu SYN v8, viz pravidlo 3a. v tabulce 2 níže.

5. Jestliže KWIC je **[word="[[:upper:]].*slava"& tag="NN(MS2|MS4|FS1).*"]**
pak v případě, že levý kontext <-2,-1> je [word="[[:upper:]].*[ui]"][lc="a"]
KWIC je **[word="[[:upper:]].*slava"]**
a pravý kontext <1,1> je [word="[[:upper:]].*ovy"]
pak velmi pravděpodobně[31] platí, že **[word="[[:upper:]].*slava"& tag="NNMS4.*" & lemma="[[:upper:]].*slav"]**
Tomuto pravidlu odporují výsledky desambiguace tvaru propria zakončeného na *-slava* v korpusu SYN v8 v 25 případech, viz pravidlo 5. v tabulce 2 níže.

5a. Jestliže KWIC je **[word="[[:upper:]].*slava"& tag="NN(MS2|MS4|FS1).*"]**
pak v případě, že levý kontext <-3,-1> je
[lc=".*rodiče|.*manžele"][word="[[:upper:]].*[ui]"][lc="a"]
KWIC je **[word="[[:upper:]].*slava"]**
a pravý kontext <1,1> je [word!="[[:upper:]].*ovy"]
pak velmi pravděpodobně platí, že **[word="[[:upper:]].*slava"& tag="NNMS4.*" & lemma="[[:upper:]].*slav"]**
Tomuto pravidlu odporují výsledky desambiguace tvaru propria zakončeného na *-slava* v korpusu SYN v8 v 27 případech, viz pravidlo 5a. v tabulce 2 níže.

6. Jestliže KWIC je **[word="[[:upper:]].*slava"& tag="NN(MS2|MS4|FS1).*"]**
pak v případě, že pravý kontext <1,3> je [lc="a"] [word="[[:upper:]].*[ui]"]
[word="[[:upper:]].*ovy"]
pak velmi pravděpodobně platí, že **[word="[[:upper:]].*slava"& tag="NNMS4.*" & lemma="[[:upper:]].*slav"]**
Tomuto pravidlu odporují výsledky desambiguace tvaru propria zakončeného na *-slava* v korpusu SYN v8 ve 4 případech, viz pravidlo 6. v tabulce 2 níže.

6a.
Jestliže KWIC je **[word="[[:upper:]].*slava"& tag="NN(MS2|MS4|FS1).*"]**

---

[31] Pravidla 5, 6 mohou generovat false positives v případě, že se v kontextu tvaru na *slava* bude vyskytovat koordinované femininum nesklonné na *i/u* (např. *Noemi, Lulu,* popř. hypokoristika *Veru, Maru, Kristi, ...*). V takovém případě nemusí být tvar na *slava* **[tag="NNMS4.*" & lemma="[[:upper:]].*slav"]**, ale může být také **[tag="NNFS1.*" & lemma="[[:upper:]].*slava"]**. Např.: *Noemi a Zdislava Královy, Miroslava a Lulu Baltazarovy*. Na takové případy jsme ovšem v korpusu SYN v8 nenarazili.

pak v případě, že levý kontext <-1,-1> je [lc=".*rodiče|.*manžele"]
KWIC je **[word="[[:upper:]].*slava"]**
a pravý kontext <1,3> je [lc="a"] [word="[[:upper:]].*[ui]"] [word!="[[:upper:]].*ovy"]
pak velmi pravděpodobně platí, že **[word="[[:upper:]].*slava"& tag="NNMS4.*" & lemma="[[:upper:]].*slav"]**
Tomuto pravidlu odporují výsledky desambiguace tvaru propria zakončeného na *-slava* v korpusu SYN v8 v 1 případě, viz pravidlo 6a. v tabulce 2 níže.
Tabulka 2

Pra-vidlo	dotaz na doklady odporující pravidlu	počet chyb (SYN v8)	příklad chyby v desambiguci (SYN v8), který je porušením pravidla
1.	[word="[[:upper:]].*"] [lc="a"] [word="[[:upper:]].*slav"& tag!="..MS1.*"] [word="[[:upper:]].*ovi"]	1417 chyb	*Profesionální pěstouni Jana a **Jaroslav** /**Jaroslav/NNFS1-----A-----** Toušovi.*
1a.	[lc=".*rodiče|.*manželé"][word="[[:upper:]].*"][lc="a"][word="[[:upper:]].*slav"& tag!="..MS1.*"] [word!="[[:upper:]].*ovi"]	212 chyb	*Manželé Veronika a **Jaroslav** /**Jaroslav/NNFS1-----A-----** si synka odvezou domů do Poličky.*
2.	[word="[[:upper:]].*slav" & tag!="..MS1.*"] [lc="a"] [word="[[:upper:]].*"] [word="[[:upper:]].*ovi"]	137 chyb	*… žili **Jaroslav**/**Jaroslav/NNFS1-----A-----** a Dana Stodolovi necelých pět let v pronájmu ...*
2a.	[lc=".*rodiče|.*manželé"][word="[[:upper:]].*slav" & tag!="..MS1.*"] [lc="a"] [word="[[:upper:]].*"] [word!="[[:upper:]].*ovi"]	1 chyba	*… bez prostředků ocitli i manželé **Dobroslav**/**Dobroslava/NNFP2-----A-----** a Kateřina Novotní.*
3.	[word="[[:upper:]].*slava" & tag!="..FS1.*"][lc="a"][word="[[:upper:]].*[^aeěiouyáéóůý]"][word="[[:upper:]].*ovi"]	139 chyb	*Vystavují **Jaroslava** /**Jaroslav/NNMS4-----A-----** a Jan Solovjevovi ...*
3a.	[lc=".*rodiče|.*manželé"][word="[[:upper:]].*slava" & tag!="..FS1.*"][lc="a"][word="[[:upper:]].*[^aeěiouyáéóůý]"][word!="[[:upper:]].*ovi"]	65 chyb	*Domů do Chrustenic si synka odvezou manželé **Drahoslava** /**Drahoslav/NNMS2-----A-----** a Miroslav.*
4.	[word="[[:upper:]].*[^aeěiouyáéóůý]"] [lc="a"][word="[[:upper:]].*slava" & tag!="..FS1.*"][word="[[:upper:]].*ovi"]	21 chyb/ FP/1[32]	*… Zasloužili se o to manželé Vítězslav a **Radoslava** /**Radoslav/NNMS2-----A-----** Doubravovi … uviděla své rodiče, Karin a **Rastislava** /**Rastislav/NNMS4-----A-----** **Jelínkovi***
4a.	[lc=".*rodiče|.*manželé"][word="[[:upper:]].*[^aeěiouyáéóůý]"] [lc="a"][word="[[:upper:]].*slava" & tag!="..FS1.*"][word!="[[:upper:]].*ovi"]	0 chyb	----------[33]

[32] Případ false positive není důsledkem restriktivní formulace pravidel, která počítají pouze s maskuliny antroponymy, jejichž nominativy singuláru končí na souhlásku, a s femininy antroponymy, jejichž akuzativ končí na [ui], tedy nikoliv s nesklonnými femininy končícími na souhlásku, či jiný vokál (*Ráchel, Karin, Rút, Ellen, Ester, Dagmar, ...*). Příčinou selhání pravidla je chyba ve shodě. Pokud by v dokladu nebyla gramatická chyba a zněl by *… uviděl své rodiče, **Karin a Rastislava Jelínkovy** …*, pravidlo by jej ignorovalo.
[33] V korpusu Araneum Bohemicum Maximum jsme vyhledali jeden doklad, který ukazuje na chybu v desambiguaci (*... Posledními majiteli byli manželé Bohumír a **Bohuslava/Bohuslav/NNMS2-----A----** Škodovy ...*). Pravidlo by chybu opravilo, přestože je v kontextu gramatická chyba, neboť tvar, ve kterém je chyba, není zahrnut do podmínky.

5.	[word="[[:upper:]].*[ui"] [lc="a"][word="[[:upper:]].*slava"& tag!="..MS4.*"][word="[[:upper:]].*ovy"]	25 chyb	*A jak nepochválit zkušené choreografy Jitku a* ***Ladislava*** */Ladislava/NNFS1-----A-----* *Košíkovy*
5a.	[lc=".*rodiče|.*manžele"][word="[[:upper:]].*[ui"][lc="a"][word="[[:upper:]].*slava"& tag!="..MS4.*"][word!="[[:upper:]].*ovy"]	27 chyb	*Eliška Kaplanová je první radostí pro rodiče Vendulu a* ***Jaroslava*** */Jaroslava/NNFS1-----A-----* *z Rájce u Chocně.*
6.	[word="[[:upper:]].*slava"& tag!="..MS4.*"][lc="a"][word="[[:upper:]].*[ui]"][word="[[:upper:]].*ovy"]	4 chyby	*… dopravila eskorta* ***Jaroslava*** */Jaroslava/NNFS1-----A-----* *a Danu Stodolovy ...*
6a.	[lc=".*rodiče|.*manžele"][word="[[:upper:]].*slava"& tag!="..MS4.*"][lc="a"][word="[[:upper:]].*[ui]"][word!="[[:upper:]].*ovy"]	1 chyba	*Měli jsme u nás manžele* ***Jaroslava*** */Jaroslav/NNMS2-----A-----* *a Aničku Lehkých z Hrobčic.*

Ověření možnosti pravidel desambiguace proprií zakončených na *-slav* v kontextu vybraných substantiv označujících místo/sídlo

V rámci hledání způsobů zlepšení desambiguace jsme se snažili zmapovat situaci spojení lemmat označujících místo/sídlo (např. *obec, město, lokalita, vesnice, městys, oblast, zámek, ves, vesnička, městečko, víska, tvrz, hrad*) a tvarů zakončených na *-slav*. V mnoha případech jsme nalezli správně desambiguovaná maskulina, neboť před lemmaty s významem místo předcházela substantiva typu *starosta* a po tvaru končícím na *-slav* následovalo příjmení (např. *starosta obce* ***Miroslav*** *Kovář*). Na základě pozorování jsme přistoupili k formulaci následujícího pravidla.

7. Jestliže KWIC je **[word="[[:upper:]].*slav"& tag="NN(MS1|FS1|FS4).*"]**
pak v případě, že levý kontext <-1,-1> je
[lemma="obec|město|lokalita|vesnice|městys|oblast|zámek|ves|vesnička|městečko|víska|tvrz|hrad"]
KWIC je **[word="[[:upper:]].*slav"]**
a pravý kontext <1,1> je [word!="[[:upper:]].*"]
pak zároveň pravý kontext <1,2> není [lemma="z"][word="[[:upper:]].*"][34],
pak platí, že **[word="[[:upper:]].*slav"& tag="NNFS1.*" & lemma="[[:upper:]].*slav"]**
Tomuto pravidlu odporují výsledky desambiguace tvaru propria oikonyma zakončeného na *-slav* v korpusu SYN v8 ve 2370 případech, viz pravidlo 7. v tabulce 3 níže.

Ověření možnosti pravidel desambiguace proprií zakončených na *-slav* v kontextu substantiv označujících funkce (předseda, trenér, mluvčí, ...)

Sledovali jsme desambiguaci tvarů zakončených na *-slav/slava* ve spojení s názvy funkcí. Na základě pozorování dat jsme navrhli následující pravidlo.

8. Jestliže KWIC je **[word="[[:upper:]].*slav"& tag="NN(MS1|FS1|FS4).*"]**
pak v případě, že levý kontext <-1,-1> je [lemma=".*starosta|.*předseda|trenér|mluvčí"]
KWIC je **[word="[[:upper:]].*slav"]**
a pravý kontext <1,1> je [word="[[:upper:]].*"]

[34] Podmínka vyloučí případy dokladů jako *… Domnělý zakladatel hradu* ***Jaroslav*** */Jaroslav/NNMS1-----A-----* *z Turnova měl sídlo věnovat …, … přijede na zámek* ***Vratislav*** */Vratislav/NNMS1-----A-----* *z Pernštejna …, … Statui věnoval obci* ***Jaroslav*** */Jaroslav/NNMS1-----A-----* *ze Šternberka ...*

pak platí, že **[word="[[:upper:]].*slav"& tag="NNMS1.*" & lemma="[[:upper:]].*slav"]** Tomuto pravidlu odporují výsledky desambiguace tvaru propria zakončeného na *-slav* v korpusu SYN v8 ve 166 případech, viz pravidlo 8. v tabulce 3 níže.[35]

Řada antroponym a řada oikonym jako další případy koordinace

Testovali jsme také pravidla na vylepšení desambiguace v případech, kdy proprium (antroponymum nebo oikonymum) zakončené na *-slav/-slava* stojí v kontextu dalších proprií. Takové řady lze definovat jako slovní tvary s počátečním velkým písmenem oddělené čárkami.

9a. Pokud v kontextu <-2,-2> a <2,2> od KWIC zakončeného na *-slav* stojí tvar s počátečním velkým písmenem označkovaný jako maskulinum neživotné, femininum[36], či neutrum oddělený od KWIC zakončeným na *-slav* čárkami, jde s velkou mírou pravděpodobnosti[37] o KWIC, které je oikonymum, femininum zakončené na -slav.

Případ false positives v řadě antroponym, v nichž stojí maskulinum životné zakončené na *-slav* v kontextu antroponym feminin, by bylo možné odstranit, pokud bychom při formulaci pravidel využili informace o příslušnosti k různým typům proprií, které zachycuje slovník *MorfFlex*.[38] Pokud v kontextu <-2,-2> a <2,2> od KWIC zakončeného na *-slav* stojí tvary s počátečním velkým písmenem označkované jako feminina_geografické názvy oddělené od KWIC zakončeného na *-slav* čárkami, jde o oikonymum femininum končící na *-slav*.[39] Pokud v kontextu <-2,-2> a <2,2> od KWIC zakončeného na *-slav* stojí tvary s počátečním velkým písmenem označkované jako feminina_jména osob oddělené od KWIC končícího na *-slav* čárkami, jde o antroponymum maskulinum životné zakončené na *-slav.*[40]

9b. Pokud v kontextu <-2,-2> a <2,2> od KWIC zakončeného na *-slav* stojí tvary s počátečním velkým písmenem označkované jako maskulina životná oddělená od KWIC zakončeného na *-slav* čárkami, jde s velkou mírou pravděpodobnosti o antroponymum

---

[35] O nedostatečném pokrytí slovníku svědčí např. interpretace tvarů *Ludoslav/X.*, Ljuboslav/X.*, Vladyslav/X.**, atd. V ostatních případech jsou tvary nerozpoznané morfologickou analýzou (značkované jako *X.**) překlepy a spojená slova. Přehled o nerozpoznaných tvarech a překlepech je možné získat tak, že pomocí p-filtru vyhledáme [tag="X.*"] v intervalu <0,0>.

[36] To, že do podmínky jsou zahrnuta feminina, je příčinou, že se mohou objevit případy false positives, neboť feminina mohou být nejen oikonyma, ale i antroponyma.

[37] Toto pravidlo neplatí 100%. První příčinou false positives je homonymie mezi apelativy a proprii (*… , 155 s . **Zahrádka, Miroslav /Miroslav/NNMS1-----A----- , Dogmata** a živý literární proces : …, … Čs. obec sokolská , 1926 **ŠLAJER , Jaroslav /Jaroslav/NNMS1-----A----- , Husitské** revoluční hnutí a husitská tradice …*). V obou případech je ve slovníku morfologického analyzátoru substantivum *Zahrádka* a Šlajer interpretováno jednak jako maskulinum životné – apelativum, jednak jako femininum – *zahrádka* nebo maskulinum neživotné – *šlajer.* Tyto interpretace byly chybně desambiguovány. Druhou příčinou false positives je fakt, že feminina antroponyma mohou stát v koordinované řadě s maskuliny antroponymy (*… narodili Justýna , Jakub , **Anna , Miroslav /Miroslav/NNMS1-----A----- , Tereza**, Adam a Kristýna ….*). Ve sledovaných datech jsou tyto případy v menšině, nicméně se objevují.

[38] Jak jsme uvedli výše, slovník MorfFlex (Hajič – Hlaváčová, 2016) zahrnuje informace o vlastních jménech. Tato rozlišení by bylo možné aplikovat v pravidlech pro desambiguaci antroponym a oikonym v koordinovaných řadách, a sice přidáním příslušné informace ve formě podmínky. Problém spočívá v tom, že výsledky užití pravidel, která využívají výsledky automatické morfologické analýzy na ní jsou z principu závislé (např. v případě řady homonymních proprií jako … *Vladislav, Vratislav, Soběslav …* by uvažované pravidlo nefungovalo).

[39] Např.: … *Luka nad Jihlavou, **Předboř, Svatoslav, Příseka** či Bítovčice ...*

[40] Např.: … *Lucie, **Andrea, Jaroslav, Tereza** a Marek ...*

maskulinum životné. False positives, které kombinují oba typy proprií, se mohou vyskytnout například v textech adres.[41]

Tato pravidla samozřejmě nelze aplikovat stejně dobře jako pravidla uváděná výše, protože se opírají o výsledky automatické morfologické analýzy (správné značkování i desambiguace tvarů v kontextu desambiguovaných tvarů končících na *-slav*[42]).

Uvedeným pravidlům odporují výsledky desambiguace tvarů proprií zakončených na *-slav* v korpusu SYN v8. Počty true positives a false positives jsou uvedeny v tabulce 3 níže, viz pravidlo 9a a 9b.

Tabulka 3

Pra-vidlo	dotaz na doklady odporující pravidlu	počet chyb (SYN v8)	příklad chyby v desambiguci (syn v8), který je porušením pravidla
7.	[lemma="obec|město|lokalita|vesnice|městys|oblast|zámek|ves|vesnička|městečko|víska|tvrz|hrad"][word="[[:upper:]].*slav" & tag!="..F.*"][word!="[[:upper:]].*"] N-filtr <0,1> [lemma="z"][word="[[:upper:]].*"]	2370 chyb	*… který se v minulých dnech v obci* ***Jaroslav /Jaroslav/NNMS1-----A-----*** *vloupal do plechové haly v areálu bývalého JZD ...*
8.	[lemma=".*starosta|.*předseda|trenér|mluvčí"][word="[[:upper:]].*slav" & tag!="..M.*"][word="[[:upper:]].*"]	166 chyb	*… řekl jejich mluvčí* ***Jaroslav /Jaroslav/NNFS4-----A-----*** *Haid ...*
9a.	[word="\,"][word="[[:upper:]].*slav" & tag!="..F.*"][word="\,"] p-filtr<-1,-1> [word="[[:upper:]].*" & tag="..[IFN].*"] p-filtr <1,1> [word="[[:upper:]].*" & tag="..[IFN].*"]	194 chyb/ FP 16	*Trasa povede přes Rakšice, Bohutice,* ***Miroslav /Miroslav/NNMS1-----A-----*** *, Dolenice, Hrušovany nad Jevišovkou, Božice a Hodonice.* *... David, Monika,* ***Miroslav /Miroslav/NNMS1-----A-----,*** *Denisa ani Tomáš ...*
9b.	word="\,"][word="[[:upper:]].*slav" & tag!="..F.*"][word="\,"] p-filtr<-1,-1> [word="[[:upper:]].*" & tag="..M.*"] p-filtr <1,1> [word="[[:upper:]].*" & tag="..M.*"]	2 chyby/ FP 1	*… i bizarně seskládaná slovanská křestní jména: Pravoslav,* ***Sudislav /Sudislav/NNFS1-----A-----*** *, Svatopluk, Slavomír, ...* *… František Nejedlo,* ***Čáslav/ Čáslav/NNFS1-----A-----,*** *Vladimír Myslivec, Semteš*

Frekvenční analýza chyb v desambiguaci

V tabulce 4 uvádíme souhrnné počty získané testováním chyb, které by bylo možné odstranit navrženými pravidly a počty false positives, které pravidla generují.

Tabulka 4

[41] Např.: *… František Nejedlo* ***, Čáslav /Čáslav/NNFS1-----A----- ,*** *Vladimír Myslivec , Semteš 2 …* Substantivum *Nejedlo* je označkováno jako maskulinum životné (příjmení), lemma je s velkým písmenem. Substantivum *Vladimír* je maskulinum životné (rodné jméno).

[42] Například v kontextu substantiva *Vratislav* se vyskytnou substantiva *Vladislav* a *Soběslav*. Vyloučení jedné/druhé interpretace u substantiva *Vratislav* pak bude nutně závislé na správné desambiguaci substantiv *Vladislav* a *Soběslav*. Ta je ovšem přesahuje možnosti automatické morfologické analýzy, neboť jde spíše o desambiguaci sémantické roviny jazyka.

typ chyb	počet chyb v desambiguaci v SYN v8	false positives
chyby v antroponymech známých osobností (LEMUR) a v kombinaci rodné jméno a příjmení	15314	0
chyby ve víceslovných oikonymech	46	0
chyby v koordinovaných skupinách	2 049	(1)
chyby v kontextu pojmenování míst	2 370	0
chyby v kontextu pojmenování funkcí	166	0
chyby v koordinovaných řadách proprií	196	17
Celkem	**20 141**	**17+(1)**

Celkem jsme na základě výše uvedených rešerší odhalili 20 141 chyb v desambiguaci word=".*(slav|slava)", které lze navrženými pravidly uvést na pravou míru.

Závěr

Propria zakončená na *-slav/-slava* jsou častými jmény osob a míst. Jako taková se hojně vyskytují v českých korpusech.

Analýza dat zahrnutých ve slovníku *MorfFlex* ukázala poměrně velmi solidní pokrytí uvedeného typu proprií. Přesto při porovnání interpretací obsažených ve slovníku *MorfFlex* s daty v korpusu SYN v8 se ukázaly některé dílčí nedostatky. K jejich odstranění navrhujeme doplnění slovníku *MorfFlex* a) o 11 interpretací lemmat, jejichž tvary jsou chybně interpretovány automatickou morfologickou analýzou a b) o frekventovaná lemmata nerozpoznaná automatickou morfologickou analýzou, jejichž seznam uvádíme výše. Přimlouváme se také za odstranění feminina antroponyma *Čáslava* a ponechání pouze příjmení maskulina *Čáslava.*

Homonymie proprií, jejichž tvary končí na *-slav/-slava* představuje ovšem daleko vážnější problém pro automatickou morfologickou analýzu. Ačkoliv jsou její výsledky vcelku uspokojivé, při podrobnějším zkoumání korpusových dat lze zjistit poměrně značné množství chyb v desambiguaci.

Navrhli jsme proto pravidla, která by pomohla odstranit existující chyby v desambiguaci sledovaných tvarů proprií. Tato pravidla lze rozdělit do několika skupin. První skupinu tvoří přehled frekventovaných víceslovných pojmenování rodné jméno + příjmení a víceslovné názvy oikonym (*Mladá/Stará Boleslav*). Tyto jednotky navrhujeme zařadit do databáze *Lemur*. Další skupinou jsou kombinace tvarů *.*slava .*ová* (rodné jméno femininum následované příjmením femininem), které lze desambiguovat na základě jednoduchého kontextového pravidla. Následují pravidla popisující tvary na *slav/slava* v různě definovaných koordinacích a pravidla, která mají zabránit chybné interpretaci tvarů zakončených na *-slav* v kontextu lexémů, které vylučují jednu interpretaci a podporují druhou. Poslední skupinu tvoří pravidla, jimiž jsme se snažili ověřit možnost využít souvýskyt tvarů – kandidátů na kongruentní skupiny. U většiny pravidel je spolehlivost 100%.

Aplikace navržených pravidel by vylepšila stávající desambiguaci v případě 20 141 chybně desambiguovaných tvarů. Tento výsledek není nikterak statisticky významný (0,54 % všech tvarů zakončených na *-slav/-slava*), přesto svou cenu má.

Chyby v desambiguaci jsou patrným důsledkem použití statistických metod desambiguace. Domníváme se, že pravidlová desambiguace opřená o výše uvedené lingvistické intuice (koordinované skupiny v širším kontextu, lexikální obsazení kolokací) je schůdnou cestou k vylepšení morfologického značkování uvedeného typu proprií. Uvážíme-li, že korpusy rostou především velkým přílivem publicistických textů, v nichž se uvedená propria vyskytují s vysokou frekvencí, pak lze předpokládat, že investice do pravidel podporujících jejich správné interpretace nemusí být bez významu. Navíc by některá pravidla bylo možné aplikovat i na dvojice podobných typů proprií (např. *.*mil/.*mila, .*mír/.*míra, Petr/Petra, Pavel/Pavla*, …).

Bibliografie

BENKO, V.: *Araneum Bohemicum Maximum, verze 15.04*. Ústav Českého národního korpusu FF UK, Praha 2015. Dostupný z: http://www.korpus.cz

BENKO, V. (2014): Aranea: Yet Another Family of (Comparable) Web Corpora. In Sojka, P. – Horák, A. – Kopeček, I. – Pala, K. (eds), *TSD 2014, LNAI 8655*, 257–264. Springer International Publishing.

DAVID, J. – ROUS, P. (2006). *Neviditelní svědkové minulosti: místní a pomístní jména na Vysočině*. Praha: Academia.

DAVID, J. (2017): Vlastní jméno posesivní. In: KARLÍK, P. – NEKULA, M. – PLESKALOVÁ, J. (eds.): *CzechEncy – Nový encyklopedický slovník češtiny.* Dostupný z: https://www.czechency.org/slovnik/VLASTNÍ JMÉNO POSESIVNÍ

HAJIČ, J. – HLAVÁČOVÁ, J. (2016): *MorfFlex CZ 160310.* Dostupný z: http://hdl.handle.net/11234/1-1673. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

HNÁTKOVÁ, M. a kol. (2014): The SYN-series corpora of written Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 160–164. Reykjavík: ELRA. ISBN 978-2-9517408-8-4. http://www.lrec-conf.org/proceedings/lrec2014/pdf/294_Paper.pdf

JELÍNEK, T. a kol. (2018): Variabilita českých frazémů v úzu. *Časopis pro moderní filologii,* Karlova univerzita, roč. 100, 2018, č. 2, s. 151−175.

KOPLOTOVCE (2020) Dostupné z: https://www.koplotovce.sk/download_file_f.php?id=621919

KŘEN, M. a kol. (2019): Korpus SYN, verze 8 z 12. 12. 2019. Ústav Českého národního korpusu FF UK, Praha 2019. Dostupný z: https://www.korpus.cz

LAMPRECHT, A. – ŠLOSAR, D. – BAUER, J. (1986): *Historická mluvnice češtiny*. Praha: Státní pedagogické nakladatelství.

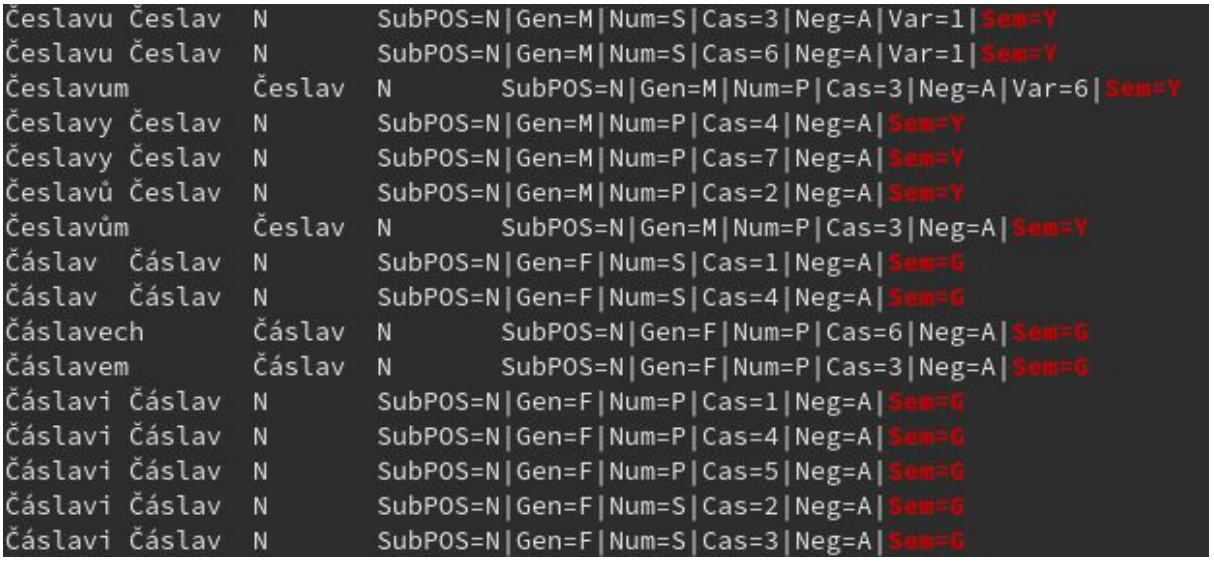
LUTTERER, I. – ŠRÁMEK, R. (2004): Zeměpisná jména v Čechách, na Moravě a ve Slezsku: slovník vybraných zeměpisných jmen s výkladem jejich původu a historického vývoje. Havlíčkův Brod: Tobiáš.

OSOLSOBĚ, K. – MACHALOVÁ, J. (2013): Hypokoristika z rodných jmen v Korpusu soukromé korespondence. In Hladká, Zdeňka a kol.. *Soukromá korespondence jako lingvistický pramen*. Brno: Masarykova univerzita, s. 33-59.

Klára Osolsobě
osolsobe@phil.muni.cz
Ústav českého jazyka
Filozofická fakulta
Masarykova univerzita
Arna Nováka 1
602 00 Brno

Hana Žižková
zizkova@phil.muni.cz
Ústav českého jazyka
Filozofická fakulta
Masarykova univerzita
Arna Nováka 1
602 00 Brno

Printscreen z morflexu



```
Česlavu Česlav  N       SubPOS=N|Gen=M|Num=S|Cas=3|Neg=A|Var=1|Sem=Y
Česlavu Česlav  N       SubPOS=N|Gen=M|Num=S|Cas=6|Neg=A|Var=1|Sem=Y
Česlavum        Česlav  N       SubPOS=N|Gen=M|Num=P|Cas=3|Neg=A|Var=6|Sem=Y
Česlavy Česlav  N       SubPOS=N|Gen=M|Num=P|Cas=4|Neg=A|Sem=Y
Česlavy Česlav  N       SubPOS=N|Gen=M|Num=P|Cas=7|Neg=A|Sem=Y
Česlavů Česlav  N       SubPOS=N|Gen=M|Num=P|Cas=2|Neg=A|Sem=Y
Česlavům        Česlav  N       SubPOS=N|Gen=M|Num=P|Cas=3|Neg=A|Sem=Y
Čáslav  Čáslav  N       SubPOS=N|Gen=F|Num=S|Cas=1|Neg=A|Sem=G
Čáslav  Čáslav  N       SubPOS=N|Gen=F|Num=S|Cas=4|Neg=A|Sem=G
Čáslavech       Čáslav  N       SubPOS=N|Gen=F|Num=P|Cas=6|Neg=A|Sem=G
Čáslavem        Čáslav  N       SubPOS=N|Gen=F|Num=P|Cas=3|Neg=A|Sem=G
Čáslavi Čáslav  N       SubPOS=N|Gen=F|Num=P|Cas=1|Neg=A|Sem=G
Čáslavi Čáslav  N       SubPOS=N|Gen=F|Num=P|Cas=4|Neg=A|Sem=G
Čáslavi Čáslav  N       SubPOS=N|Gen=F|Num=P|Cas=5|Neg=A|Sem=G
Čáslavi Čáslav  N       SubPOS=N|Gen=F|Num=S|Cas=2|Neg=A|Sem=G
Čáslavi Čáslav  N       SubPOS=N|Gen=F|Num=S|Cas=3|Neg=A|Sem=G
```

Miroslavové	Miroslav	N	SubPOS=N|Gen=M|Num=P|Cas=5|Neg=A|Sem=Y
Miroslavu	Miroslav	N	SubPOS=N|Gen=M|Num=S|Cas=3|Neg=A|Var=1|Sem=Y
Miroslavu	Miroslav	N	SubPOS=N|Gen=M|Num=S|Cas=6|Neg=A|Var=1|Sem=Y
Miroslavum	Miroslav	N	SubPOS=N|Gen=M|Num=P|Cas=3|Neg=A|Var=6|Sem=Y
Miroslavy	Miroslav	N	SubPOS=N|Gen=M|Num=P|Cas=4|Neg=A|Sem=Y
Miroslavy	Miroslav	N	SubPOS=N|Gen=M|Num=P|Cas=7|Neg=A|Sem=Y
Miroslaví	Miroslav	N	SubPOS=N|Gen=F|Num=P|Cas=2|Neg=A|Sem=G
Miroslaví	Miroslav	N	SubPOS=N|Gen=F|Num=S|Cas=7|Neg=A|Sem=G
Miroslavích	Miroslav	N	SubPOS=N|Gen=F|Num=P|Cas=6|Neg=A|Sem=G
Miroslavím	Miroslav	N	SubPOS=N|Gen=F|Num=P|Cas=3|Neg=A|Sem=G
Miroslavů	Miroslav	N	SubPOS=N|Gen=M|Num=P|Cas=2|Neg=A|Sem=Y
Miroslavům	Miroslav	N	SubPOS=N|Gen=M|Num=P|Cas=3|Neg=A|Sem=Y

Boleslav	Boleslav	N	SubPOS=N|Gen=F|Num=S|Cas=1|Neg=A|Sem=G
Boleslav	Boleslav	N	SubPOS=N|Gen=F|Num=S|Cas=4|Neg=A|Sem=G
Boleslav	Boleslav	N	SubPOS=N|Gen=M|Num=S|Cas=1|Neg=A|Sem=Y
Boleslava	Boleslav	N	SubPOS=N|Gen=M|Num=S|Cas=2|Neg=A|Sem=Y
Boleslava	Boleslav	N	SubPOS=N|Gen=M|Num=S|Cas=4|Neg=A|Sem=Y
Boleslavama	Boleslav	N	SubPOS=N|Gen=M|Num=P|Cas=7|Neg=A|Var=6|Sem=Y
Boleslave	Boleslav	N	SubPOS=N|Gen=M|Num=S|Cas=5|Neg=A|Sem=Y
Boleslavech	Boleslav	N	SubPOS=N|Gen=M|Num=P|Cas=6|Neg=A|Sem=Y
Boleslavem	Boleslav	N	SubPOS=N|Gen=M|Num=S|Cas=7|Neg=A|Sem=Y
Boleslavemi	Boleslav	N	SubPOS=N|Gen=F|Num=P|Cas=7|Neg=A|Sem=G

Jaroslav	Jaroslav	N	SubPOS=N|Gen=F|Num=S|Cas=1|Neg=A|Sem=G
Jaroslav	Jaroslav	N	SubPOS=N|Gen=F|Num=S|Cas=4|Neg=A|Sem=G
Jaroslav	Jaroslav	N	SubPOS=N|Gen=M|Num=S|Cas=1|Neg=A|Sem=Y
Jaroslava	Jaroslav	N	SubPOS=N|Gen=M|Num=S|Cas=2|Neg=A|Sem=Y
Jaroslava	Jaroslav	N	SubPOS=N|Gen=M|Num=S|Cas=4|Neg=A|Sem=Y
Jaroslavama	Jaroslav	N	SubPOS=N|Gen=M|Num=P|Cas=7|Neg=A|Var=6|Sem=Y
Jaroslave	Jaroslav	N	SubPOS=N|Gen=M|Num=S|Cas=5|Neg=A|Sem=Y
Jaroslavech	Jaroslav	N	SubPOS=N|Gen=F|Num=P|Cas=6|Neg=A|Sem=G
Jaroslavech	Jaroslav	N	SubPOS=N|Gen=M|Num=P|Cas=6|Neg=A|Sem=Y
Jaroslavem	Jaroslav	N	SubPOS=N|Gen=F|Num=P|Cas=3|Neg=A|Sem=G
Jaroslavem	Jaroslav	N	SubPOS=N|Gen=M|Num=S|Cas=7|Neg=A|Sem=Y
Jaroslavi	Jaroslav	N	SubPOS=N|Gen=F|Num=P|Cas=1|Neg=A|Sem=G