Příloha B

1 Popis seznamu lemmat pro generování tvarů

Seznam lemmat představuje základ pro automatické generování tvarů – lze z něj získat seznam tvarotvorných základů, které po kombinaci s koncovkami nebo zakončeními příslušných vzorů a po aplikaci alternací umožní vytvoření tvarů celého paradigmatu. V seznamu jsou lemmata roztříděna k jednotlivým vzorům, které byly představeny v kapitole 2, a jsou k nim přiřazeny také alternace, tedy změny tvarotvorného základu, na které není možné usoudit z hláskové podoby lemmatu, nebo jejichž odvození pomocí pravidel by bylo velmi komplikované. Alternace byly systematicky popsány v kapitole 3. Další podoby takto získaných tvarů mohou být získány aplikací formálních změn popsaných v Příloze A.

Celkově seznam obsahuje asi 29 000 lemmat přiřazených asi ke 100 vzorům. Různých značek pro alternace bylo přiřazeno asi 120.

V této kapitole se nejprve stručně popisují jednotlivé fáze vzniku seznamu, poté se komentují problematické jevy, které se při jeho vytváření objevily, a nakonec je uvedena ukázka seznamu. Celý seznam je k dispozici jako Příloha C. Oproti seznamu lemmat, který byl přílohou disertační práce, je seznam v jednotlivostech upraven (několik lemmat bylo vyškrtnuto, protože se jednalo o nástupnické podoby lemmat uvedených v jiných slovnících v podobě k roku 1300, nebo protože se ukázalo, že jsou to lemmata s cizí deklinací, u několika lemmat bylo potřeba pozměnit značku alternace, celkově se jednalo asi o 40 lemmat). Tyto změny vyplynuly z prvních pokusů s automatickým generováním tvarů.

1.1 Postup vytváření seznamu lemmat

Seznam lemmat pro generování tvarů byl vytvářen pomocí nástroje pro práci s komplikovanými daty OpenRefine, který umožňuje složité filtrování a dotazování se na obsah buněk pomocí regulárních výrazů.

Seznam vznikl v několika krocích. Prvním byla automatická extrakce apelativních lemmat ze staročeských slovníků Vokabuláře webového a jejich přiřazení k deklinačnímu typu pomocí kombinace rodu, koncovky v reprezentativním tvaru a opěrného tvaru. Tento krok provedl Boris Lehečka na základě tabulky kombinací koncovek v reprezentativním tvaru a opěrných tvarů, kterou jsem připravila.

Pro extrakci byl zvolen tento postup: U hesel ze všech staročeských slovníků byly sjednoceny uváděné informace o jmenném rodu a morfologické charakteristice (např. f × ž, -a, -u × -a/-u). Signálem pro převzetí heslového slova do seznamu jako apelativa bylo uvedení jmenného rodu, velikost prvního písmene a jednoslovnost hesla. Poté byla ke každému heslovému slovu vytvořena identifikace na základě kombinace heslo – jmenný rod – opěrný tvar GEN.SG (GEN.PL). Pokud mělo více hesel stejnou identifikaci, bylo v seznamu ponecháno pouze heslo ze StčS. Pokud mezi hesly s touž identifikací takové neexistovalo, bylo vybráno heslo z ESSČ a v hierarchii následovala hesla z GbSlov a MSS. Tato hierarchie odráží míru propracovanosti a sjednoceného uvádění informací v daných slovnících. Všechna hesla z MSS, která nemají uveden opěrný tvar, byla odstraněna ze seznamu, protože neumožňovala přiřazení vzoru a očekává se jejich nahrazení hesly z ESSČ. Pokud mělo dané heslové slovo uvedeno více opěrných tvarů a jeden rod, znásobilo se jeho uvedení v seznamu podle počtu opěrných tvarů. Vzory byly heslům přiřazeny na základě zmíněné tabulky kombinací, automatická procedura signalizovala, pokud bylo přiřazení vzoru nejisté, nebo nemožné (např. neodpovídalo si zakončení lemmatu a opěrný tvar). U mužských o- a jo-kmenů, kde je vzor ovlivněn významem daného jména, automatická procedura brala v potaz první význam heslového slova (např. při výskytu slova *člověk* v popisu přiřazovala vzory pro osoby).

V dalších fázích jsem již pracovala samostatně. Nejprve byla ze seznamu odstraněna všechna hesla, která automatická procedura na základě rodu vyhodnotila jako apelativa, ale jednalo se o zájmena nebo číslovky. Dále byla ze seznamu odstraněna všechna apelativa s cizí deklinací a skloňovaná podle adjektivní deklinace. Ve všech případech, kde byla automaticky vytvořena neexistující kombinace koncovky lemmatu a koncovky/zakončení opěrného tvaru (v případě více lemmat v rámci jednoho hesla), bylo přiřazení vzoru provedeno ručně (automatická procedura nechávala vzor nerozhodnutý). Celkem se tato první korekce týkala asi 6 000 lemmat.

Třetím krokem bylo ruční přiřazení apelativ ke vzorům v těch případech, kdy deklinační typ obsahuje více než jeden vzor (např. pro mužské o-kmeny bylo třeba ručně roztřídit jména osob, živočichů a neživých entit, u ženských ja-kmenů nalézt všechna jednoslabičná apelativa apod.). U deklinací, kde se na vzor usuzuje na základě významu, bylo toto přiřazení provedeno na základě prvního významu, který byl jako jediný extrahován automatickým předzpracováním. Obsahuje-li tedy popis významu zároveň význam ‚osoba' i ‚neživá věc' a význam ‚osoba' je řazen na prvním místě, je apelativum v seznamu lemmat

zařazeno pouze u jmen osob. Tohoto problému jsem si vědoma, ale ruční procházení tisíců heslových statí, které by ověřilo, zda se v nich neobjevuje více významů, bylo mimo možnosti této práce. U apelativ pojmenovávajících bájné bytosti jsem při přiřazování ke jménům osob a živočichů přihlížela k tomu, zda se jedná o bytost spíše lidského nebo spíše zvířecího charakteru.

V dalším kroku byla označena singularia a pluralia tantum (příklad viz níže).

Posledním krokem bylo ruční kódování alternací vázaných na lemma.

1.2 Problematické jevy

1.2.1 Nepůvodní podoby lemmatu

Při řazení nepůvodních podob lemmat ke vzorům je využívána historická perspektiva, lemma patří tam, kam patří jeho tvarotvorný základ – lemmata *dci* i *dcera* jsou řazena k r-kmenům, *sáni* i *sáně* k i-kmenům, *bukev*, *bukva* i *bukvě* k ъv-kmenům atd. U lemmat, kde toto kritérium nelze aplikovat, jsou zaváděny nové vzory (to se týká především lemmat, která nelze řadit historicky ani k ja-kmenům, ani k i-kmenům – viz vzor obnož u ja-kmenů, část 2.13.2.1).

1.2.2 Duplicity

Komentář vyžaduje také zacházení s duplicitami, tedy lemmaty, která se vyskytla ve více staročeských slovnících zároveň. Automatická extrakce duplicity do seznamu zaváděla jen jednou v případě, že v různých slovnících u nich byla uvedena stejná charakteristika (heslové slovo, rod, tentýž opěrný tvar).

V ostatních případech se lemmata uvedená ve více staročeských slovnících zároveň dostala do seznamu pro generování vícekrát (např. v případech, kdy různé slovníky uvádějí stejný opěrný tvar, ale v různém formátu: např. u lemmatu *aksamit* uvádí ESSČ uvádí opěrný tvar s koncovkami -*a*/-*u*, v GbSlov je uvedeno jen -*a*). Použitý software dovoluje duplicity automaticky najít, a tedy i vyfiltrovat (celkem bylo nalezeno asi 1 900 duplicitních lemmat). Po vyfiltrování následovala kontrola, zda se takto ze seznamu neztratily případy, ve kterých jsou sice lemmata formálně totožná, ale patří k jiným deklinacím. Např. výraz *podběl* je jednak o-kmenová maskulinum, jednak i-kmenové femininum, nebo výraz *pikús* je v MSS přiřazen k jo-kmenům, ve StčS k o-kmenům. Taková lemmata byla znovu zařazena do seznamu pro generování tvarů a obě interpretace byly zachovány (jednalo se asi o 200 lemmat). Naopak homonyma jako *bor* (‚*zástup*', ‚*les*') nebo *koba* (‚*přání*', ‚*havran*') byla v seznamu pro generování ponechána pouze jednou, protože obě homonyma patří ke stejnému vzoru, a pro tvoření tvarů tak jsou pokryty všechny možnosti.

1.2.3 Propojování lemmat

Propojování lemmat – tedy zavádění takových informací (přiřazení ke vzoru nebo uvedení alternací tvarotvorného základu), které umožní tvořit i tvary, které mohou patřit k jiným lemmatům (kromě případů náhodné homonymie) – je v seznamu pro generování tvarů omezeno v souladu s principy představenými u úvodu této práce na dvě situace. Týká se jednak propojení tvarů zástupců malých deklinací (např. lemmata jako *dci* a *dcera* jsou zařazena k r-kmenům a ve vzoru jsou tvary zachyceny tak, aby po zadání lemmatu *dci* bylo možné nalézt i tvar *dcera* a naopak), jednak propojení tvarů lemmat lišících se v NOM.SG pouze vkladným *e* (např. *bázn*/*bázen*) a dále případů výjimečných, kde by bylo přiřazování tvarů k jednotlivým lemmatům velmi komplikované (např. u apelativa *dska*/*cka*/*deska* atd. nebo u apelativa *dřvi*/*dveři*/*dvéře* atd., viz části 3.3.4 a 3.3.6). Odůvodnění a diskusi viz v částech 2.16 (ъv-kmeny), 2.17 (r-kmeny) a 2.21 (střední n-kmeny), u alternací v částech 3.1 a 3.2.1.

1.2.4 Doplnění a opravy morfologických charakteristik

U apelativ, která neměla ve staročeských slovnících přiřazen rod a/nebo opěrný tvar, byl vzor přiřazen podle apelativ s podobnými vlastnostmi. Např. lemma *prohlub* má v StčS přiřazeny koncovky *-a/u*, ale rod není uveden, na základě podobných apelativ byl k němu přiřazen mužský rod, podobně u apelativ *vzróst*, *zróst* a *zrost*, která nemají uveden ani rod ani opěrný tvar, atd.

Stejně jsem postupovala i u lemmat, kde byla kombinace opěrného tvaru a rodu uvedena nejspíše chybně – např. apelativum *hřebenářstvo* bylo ve verzi ESSČ, se kterou jsem pracovala, uvedeno jako femininum, *ztynk* jako neutrum, *sedlce* jako maskulinum, *mužnost* v MSS jako maskulinum, *známek* jako neutrum, StčS uvádí *plázě* (*-ěte*) jako maskulinum atd.

1.2.5 Oprava podoby lemmatu

U některých lemmat bylo třeba opravit podobu lemmatu – např. lemma *přímluvčie* uváděné v MSS má pro staročeské období náležitou podobu *přímluvčí*, *vřěseň* má náležitou podobu *vřesen* nebo *jiedce* podobu *jiedcě*. V lemmatu uváděném ve StčS byla opravena podoba *prosřědnice* na *prosřědnicě*. Opravována byla jen lemmata s chybou výslovně zmíněnou v ESSČ nebo jinak nezpochybnitelnou.

1.2.6 Doplnění lemmatu

V souladu s principy popsanými v úvodu práce byla lemmata do seznamu pro generování tvarů doplňována jen výjimečně. Jednalo se o dvě základní situace. Za prvé byla doplňována lemmata, která byla v některém ze slovníků uvedena s vícenásobným opěrným tvarem i rodem. Taková lemmata se automatickou procedurou dostala do seznamu jen jednou (s nerozhodnutým vzorem) a v seznamu lemmat pro ně bylo nutné vytvořit další řádky pro další morfologické interpretace (např. lemma *žól* je obsaženo pouze v ESSČ a je mu přiřazena charakteristika ženského i-kmenu a mužského o-kmenu,

v seznamu tedy bylo potřeba vytvořit druhý řádek pro druhou charakteristiku). Za druhé byla lemmata doplňována v případě, že tvary vyskytující se v textech dané lemma opravňují, ale ve staročeských slovnících se buď dané lemma nevyskytuje (např. v MSS je uvedeno pouze lemma *tróska*, ale doklady jako *vezmi trusky železné* (LékŽen) opravňují zavedení podoby *truska*), nebo se zde vyskytuje pouze ve formě odkazu k plnému heslu (např. *řiepa* viz *řěpa*). Odkazy k heslu nebyly automatickou procedurou při první fázi tvorby seznamu rozpoznány jako apelativa a zařazeny do seznamu. Lemmata doplněná ve druhém kroku byla nalezena v důsledku práce na popisu jednotlivých témat, neproběhlo žádné systematické vyhledání všech odkazů a z pochopitelných důvodů jsem se nesnažila ani nalézt všechny tvary nezařaditelné k lemmatům. Je tedy nanejvýš pravděpodobné, že v textech ITB se vyskytují další tvary, pro které staročeské slovníky neuvádějí žádné lemma, nebo jen lemma ve formě odkazu. Zpřesnění v tomto směru může však přinést až automatická morfologická analýza textů ITB, případně doplnění nových lemmat z ESSČ a všech lemmat uvedených pouze jako odkaz do seznamu pro generování tvarů.

Doplněná lemmata jsou v seznamu pro generování tvarů ve sloupci poznámka označena jako pridane_lemma_ve_sl_neni a pridane_lemma_ve_sl_je, podle toho, zda dané lemma ve staročeských slovnících zavedeno není vůbec (ve slovníku není), nebo jen ve formě, která neumožnila jejich automatické zařazení do seznamu (ve slovníku je).

1.3 Ukázka části seznamu

Tabulka 1 ukazuje část seznamu lemmat pro generování tvarů. Vzhledem k šířce stránky byl vynechán sloupec, ve kterém je pro každé lemma vypsán zdroj, ze kterého bylo lemma přejato (ESSČ, StčS, GbSlov, MSS), a sloupec s výpisem prvního významu.

Kromě lemmat obsahuje seznam sloupec se vzorem, ke kterému dané lemma patří. Značka vzoru obsahuje zkratku subst signalizující, že se jedná o vzor substantivní, dále je v jeho názvu uvedena zkratka pro rod, poté je vypsán kmen, ke kterému apelativum náleží, a nakonec se ve značce vyskytuje vzorové slovo. Sloupec s názvem omezení uvádí, zda je dané apelativum singulare, nebo plurale tantum. U takových apelativ se budou podle daného vzoru tvořit pouze tvary singuláru, nebo plurálu. Ve sloupci alternace je uvedena značka pro alternace. Značka obsahuje popis typu alternace, určení, v jakých tvarech k alternaci dochází a jaké sekvence hlásek se při alternaci ve tvaru mění. Sloupec poznámka je vyplněn pouze u doplněných lemmat – uvádí se v něm, zda doplněné lemma staročeské slovníky uvádějí, nebo je zavedeno nově. Poznámka je zavedena zejména proto, aby bylo možné doplněná lemmata v případě potřeby snadno vyfiltrovat.

lemma	vzor	omezení	alternace	poznámka
brúk	subst.m.o-kmen.pták			
brunát	subst.m.o-kmen.dub			
brunátnost	subst.f.i-kmen.kost			
bruně	subst.f.ja-kmen.dýně	plurale_tantum		
brunieř	subst.m.jo-kmen.muž			
brus	subst.m.o-kmen.dub			
brusec	subst.m.jo-kmen.hrnec		E0-bezkoncovkove_tvary-KeK/tvary_s_koncovkou-KK	
brúsek	subst.m.o-kmen.zvuk		E0-bezkoncovkove_tvary-KeK/tvary_s_koncovkou-KK	
brusič	subst.m.jo-kmen.muž			
brusidlo	subst.n.o-kmen.město		E0-bezkoncovkove_tvary-KeK+KK/tvary_s_koncovkou-KK	
brusiny	subst.f.a-kmen.žena	plurale_tantum		
brusnič	subst.f.ja-kmen.duš			
brusnička	subst.f.a-kmen.žena		E0-bezkoncovkove_tvary-KeK/tvary_s_koncovkou-KK	
brúšenie	subst.n.ьjo-kmen.znamenie			
brvař	subst.m.jo-kmen.muž			
brvno	subst.n.o-kmen.město		E0-bezkoncovkove_tvary-KeK/tvary_s_koncovkou-KK	
brykyš	subst.m.jo-kmen.meč			
bryl	subst.m.o-kmen.dub			
brynda	subst.f.a-kmen.žena		kvantita-GEN.PL-dlouha+kratka	
bryždl	subst.m.o-kmen.dub			pridane_lemma_ve_sl_je
bryždl	subst.m.jo-kmen.meč			pridane_lemma_ve_sl_je
bryždl	subst.f.i-kmen.kost			
brýždl	subst.m.o-kmen.dub			
brýždl	subst.m.jo-kmen.meč			pridane_lemma_ve_sl_je
brzěnín	subst.m.o-kmen.zeměnín		zaklad-cele_paradigma-ěnín+an+ěn	

Tabulka 1: Ukázka části seznamu pro generování tvarů (celý seznam je přiložen na CD)

Celkový přehled vzorů, které jsou k lemmatům zavedeny, s počty lemmat k nim přiřazených obsahuje Tabulka 16 uvedená ve shrnutí práce (část 4.2.2). Klíč ke značení jednotlivých alternací je uveden v úvodu kapitoly Alternace (kapitola 3). Ve shrnutí práce (část 4.3.2) je také možné nalézt Tabulku 18 shrnující nejčastější typy alternací a počty lemmat k nim přiřazených.