OPEN ACCESS

Na cestě k lemmatizaci staročeských textů: data, software, aplikace[1]

Pavlína Synková — Boris Lehečka — Ondřej Svoboda

ABSTRACT:
Towards the lemmatization of Old Czech texts: data, software, applications. This paper introduces the description of Old Czech common nouns developed and used in a tool for tagging and lemmatizing common nouns occurring in transcribed digital editions of Old Czech texts. This description consists of four parts: the first features an overview of all declension type endings (approx. 100 declension patterns), the second part analyses alternations in the morphological basis accompanying declension (approx. 120 types of alternations), the third part deals with formal changes connected mainly with the language's historical development (approx. 100 formal changes) and, finally, the fourth part contains a list of lemmas extracted from modern dictionaries of Old Czech (approx. 29 000 lemmas). Furthermore, the paper introduces the software developed and used for this purpose, namely i) the tool which makes it possible a) to generate word forms and subsequently search for multiple word forms in the texts at once, b) to create lists of word forms filtered by sequences of characters occurring at the end of the word forms, ii) the tool for assigning a declension pattern to a lemma, and iii) the tool enabling work with large databases. Finally, the paper describes two applications developed on the basis of Old Czech common noun description, i.e. i) a database of Old Czech common noun declension patterns connected with Old Czech dictionaries and the Old Czech text bank, ii) a tool for generating word forms, which is used for the lemmatization and tagging of Old Czech texts.

KLÍČOVÁ SLOVA / KEY WORDS:
apelativa, lemmatizace, NLP software a aplikace, stará čeština, tagování, XML
common nouns, lemmatization, NLP software and applications, Old Czech, tagging, XML

1. ÚVOD

Cílem článku je představit popis staročeské apelativní deklinace, jehož část vyšla knižně (Synková, 2017), přiblížit softwarové nástroje, které byly při jeho vytváření použity, a popsat aplikace, které doposud na jeho základě vznikly: přehled staročeských vzorů propojených se slovníky *Vokabuláře webového* a staročeskou textovou bankou a dále generátor tvarů, který využívá popis při tagování a lemmatizaci staročeských textů. Následující text představuje nejprve popis staročeských apelativ (část 2), poté se věnuje dříve existujícím i nově naprogramovaným softwarovým nástrojům využitým při jeho přípravě (část 3) a nakonec vyvinutým aplikacím (části 4 a 5).

Děkujeme recenzentům za jejich připomínky a návrhy na opravy. Zůstaly-li v článku nějaké chyby, pak pouze vinou autorů.

1 Vznik příspěvku byl podpořen projektem Ministerstva školství, mládeže a tělovýchovy č. LM2015081 Výzkumná infrastruktura pro diachronní bohemistiku (akronym RIDICS) v rámci Projektu velkých infrastruktur pro VaVaI.

2. FORMÁLNÍ POPIS STAROČESKÉ MORFOLOGIE

Nedávno dokončený formální popis[2] staročeské apelativní deklinace byl od počátku koncipován tak, aby mohl sloužit jako základ pro automatické vygenerování tvarů spojených s morfologickými charakteristikami a lemmatem a tyto tvary a informace mohly být poté využity pro přiřazování morfologických kategorií (rodu, čísla a pádu) a lemmatu, popř. hyperlemmatu,[3] tvarům vyskytujícím se v transkribovaných staročeských textech. Staročeským obdobím se přitom ve shodě s obecně přijatou periodizací myslí období od vzniku souvislých českých textů zhruba do roku 1500. Substantiva byla vybrána proto, že v současné češtině pokrývají zhruba 30 % textu (Bartoň et al., 2009, s. 130), tedy nejvíce ze všech slovních druhů, a není důvod předpokládat, že pro staročeské texty bude jejich podíl výrazně jiný. V celém popisu se zohledňují staročeské texty pouze v transkripci (zásady transkripce viz Černá & Lehečka, 2015; Daňhelka, 1957, 1963, 1985; Staročeský slovník, 1968). Nejrozsáhlejší soubor staročeských transkribovaných a elektronicky přístupných textů představuje v současnosti staročeská textová banka (2008–dosud) budovaná v Ústavu pro jazyk český AV ČR, v. v. i.[4]

Pro automatickou morfologickou analýzu představuje transkripce velké usnadnění, protože standardizuje písmo i pravopis, zároveň je však třeba mít na zřeteli, že každá transkripce je interpretací a je do jisté míry závislá na rozhodnutí editora textu. Tento aspekt transkribovaných textů musí mít na paměti každý badatel, který s nimi pracuje: měl by odhadnout, do jaké míry může mít transkripce na jím zkoumanou otázku vliv, a potom se buď na materiál víceméně spolehnout (např. při slovosledném zkoumání), nebo ho použít jen pro vyhledání relevantních míst a tato místa najít v původních pramenech (tj. rukopisech nebo tiscích; např. při zkoumání jotace nebo kvantity).

Pro popis staročeské apelativní deklinace byly využity historické mluvnice, staročeské texty a moderní slovníky staré češtiny. Historické mluvnice (především Gebauer, 1960 a 1963; doplňkově i Gebauer, 2007; Trávníček, 1935; Vážný, 1965; Komárek, 1969; Lamprecht, Šlosar & Bauer, 1986; Komárek, 2012) sloužily jako východisko práce, jejich tvrzení byla systematicky ověřována a doplňována pomocí textů interní verze staročeské textové banky, jejíž součástí jsou i texty, které prozatím neprošly finální redakční kontrolou (označujeme je jako nespolehlivé). Verze použitá pro většinu témat obsahovala 7,6 milionu tokenů, z nichž 3,2 mil. tokenů bylo nespolehlivých. K prohledávání této textové banky byl využit nástroj *Analýza tokenů v Excelu* (2015), který umožňuje a) po zadání tvarotvorných základů (tj. části slova, která je společná pro všechny tvary paradigmatu) a koncovek/zakončení[5] generovat tvary a hromadně

2 Za formální popis považujeme počítačově zpracovatelné údaje uložené v předem definované struktuře (v tomto případě ve formátu XML).

3 Za hyperlemma považujeme „základní tvar lexému v hláskové podobě k roku 1300“ (Černá & Lehečka, 2015, s. 67).

4 Aktuální verze má 4 929 141 tokenů (stav ke dni 1. 11. 2017, verze dat 1.1.3).

5 Jako *koncovky* (např. *-ě* u tvaru *rámě*) zde označujeme pouze takové tvarotvorné formanty (TF), z nichž nelze vyčlenit kmenotvorný sufix. Termín *zakončení* používá Synková (2017)

je hledat v textech, b) prohledávat tvary (ve smyslu typů) vyfiltrované na základě hlásek, jimiž tvar končí (více viz níže v části 3.1.1). Pokud bylo třeba použít materiál z nespolehlivých textů, byly doklady kontrolovány přímo v kopiích rukopisů nebo edic, ze kterých texty pocházejí. Slovníky pro starou češtinu zpřístupněné elektronicky ve *Vokabuláři webovém* (GbSlov: Slovník staročeský, 1970;[6] MSS: Malý staročeský slovník, 1978; StčS: Staročeský slovník, 1968–2008; ESSČ: Elektronický slovník staré češtiny, 2006–dosud) sloužily jako základ pro přehled o slovní zásobě staročeského období. Žádný z nich však nepokrývá staročeské období celé a slovníky se metodologicky liší, proto je třeba do budoucna počítat s rozšiřováním a zpřesňováním údajů.[7]

V průběhu prací se ukázalo, že popis musí tvořit čtyři základní části, aby pokryl všechny aspekty staročeské apelativní deklinace potřebné pro generování tvarů.

První část představuje popis zakončení jednotlivých deklinačních typů (odpovídajících kmenům). Zakončení jsou v ní popsána jednak v textové formě, jednak ve formě tabulek. V tabulkách se zohledňuje i původ a doložení koncovek. V rámci jednotlivých deklinačních typů je popsán různý počet vzorů. Vzor byl definován jako jedinečný soubor zakončení, kterými se tvoří tvary určité skupiny slov (např. pojmenování pro osoby nebo apelativ s tvarotvorným základem zakončeným na veláru). Pro přehlednější prezentaci vzorů bylo zvoleno rozdělení zakončení daného deklinačního typu na zakončení patřící k tzv. substrátu, který je společný pro celý deklinační typ (např. u mužských *o*-kmenů v LOC.PL je společnou koncovkou všech zástupců *-iech*), a zakončení, která mají jen někteří zástupci deklinačního typu (např. koncovku *-ech* v LOC.PL mají jen *o*-kmeny s tvarotvorným základem zakončeným na jinou hlásku než veláru). Vzor se pak skládá ze substrátu a zakončení pro své konkrétní zástupce.[8] Popis usiloval o vhodnou minimalizaci počtu vzorů, takže nebyly zaváděny samostatné vzory pro singularia a pluralia tantum[9] ani pro apelativa kolísající mezi více deklinačními typy, pokud se nepodařilo nalézt tvary odlišné od tvarů pravidelných pro daný deklinační typ. Kromě vzorů pro tradiční deklinační typy byly zavedeny i vzory pro jména nesklonná a pro apelativa výjimečná, jež nelze zařadit k pravidelným deklinačním typům a u nichž zároveň nejsou doloženy tvary, ze kterých by bylo možné paradigma sestavit. Celkem bylo popsáno 96 vzorů ve 22 deklinačních typech (nejvíce zástupců mají mužské *o*-kmeny, střední *ьjo*-kmeny a ženské *a*-kmeny).

pro TF se zřetelným kmenotvorným sufixem a koncovkou (např. tvar *rameno* obsahuje zakončení *-eno*). Pro jednoduchost zde tento termín používáme i jako zastřešující pro oba druhy TF.

6 První vydání slovníku vyšlo v letech 1903 a 1916.

7 Zejména z Elektronického slovníku staré češtiny (2006–dosud), který postupně jednotným způsobem zpracuje kompletní dochovanou staročeskou slovní zásobu (s výjimkou hesel s náslovím *n-* až *při*, která jsou zpracována ve Staročeském slovníku (1968–2008)).

8 Kromě toho byly pro lemmata, u kterých se z hlediska daného deklinačního typu vyskytují jedinečné nepravidelnosti (např. lemma *dietě* má část paradigmatu podle středních *nt*-kmenů a část podle ženských *i*-kmenů), zavedeny i tzv. vzory samostatné, u kterých se se substrátem nepočítá, je rovnou zavedeno celé paradigma.

9 Omezení paradigmatu daného apelativa pouze na tvary singuláru nebo plurálu je zavedeno v seznamu pro generování tvarů (viz níže) jako samostatná charakteristika.

U deklinačních typů, které mají do 80 zástupců a představují ve staročeském období marginální a rozpadající se typy deklinace, byl vzor stanoven po ruční analýze všech tvarů uvedených v historických mluvnicích a doložených v použité verzi textové banky. Pro ostatní (tj. z hlediska počtu zástupců větší) deklinace nebylo možné prozkoumat všechny tvary, ručně byly proto zkoumány zvláště tvary přejaté z jiných deklinací, uváděné v mluvnicích jako zvláštní, nedoložené apod., aby byl popis uváděný v mluvnicích ověřen a doplněn.

V druhé části byly popsány alternace, tedy změny tvarotvorného základu, které není možné nebo výhodné zavádět ve formě obecného pravidla pro přepis určité sekvence písmen na jinou, protože se nevyskytují u všech lemmat s danou formální stavbou (srov. *pes-ø*[10] — *ps-a*, ale *les-ø* — *les-a*; *krám-a* — *krav-ám*, ale *krás-a* — *krás-ám*) nebo by jejich zavádění formou pravidla bylo příliš složité (srov. *hvězd-a* — *hvězd-ø*, *otázk-a* — *otázek-ø*, *šacht-a* — *šacht-ø*/*šachet-ø*). Alternace jsou popsány v textové formě a pro jednotlivé typy alternací jsou zavedeny značky, jež jsou použity ve čtvrté části práce — seznamu pro generování tvarů — jako signál, jaká alternace tvarotvorného základu se u daného lemmatu objevuje. Značka se skládá z typu alternace, tvarů, ve kterých se alternace vyskytuje, a její konkrétní realizace (např. značka „Eø-bez-koncovkove_tvary-KeK/tvary_s_koncovkou-KK“ popisuje alternaci jako *pes* — *psa* nebo *otázka* — *otázek*, při níž se střídá v daném místě tvarotvorného základu *e* a nic (ø). V tvarech s nulovou koncovkou je tvarotvorný základ zakončen na sekvenci KeK, tedy konsonant (K) — samohláska *e* — konsonant (*pes*, *otázek*), v tvarech s nenulovou koncovkou končí tvarotvorný základ na sekvenci dvou konsonantů (KK, *ps-*, *otázk-*). Celkem bylo nalezeno asi 120 typů alternací, ale jen 15 z nich je natolik častých, že se vyskytují u více než 10 lemmat. Nejvíce lemmat zasahují alternace působené jerovým nebo vkladným *e*.

Ve třetí části byly popsány formální proměny psaných tvarů, které lze zavést pomocí pravidla. Jedná se jednak o hláskové změny spojené s vývojem tvarů v daném období (např. *viera* — *víra*, *bóh* — *buoh*), přičemž se vychází z hláskové podoby předpokládané k roku 1300,[11] jednak o změny vznikající při spojování tvarotvorných základů a zakončení, z nichž některé jsou jen záležitostí ortografie (např. *vlk+i* = *vlci*, *líň+em* = *líněm*). Formální změny jsou popsány ve formě textu a zároveň ve schematické formě jako pravidla, jaká písmena ubývají, přibývají či se mění na jaká (případně i v jakém kontextu). Celkem bylo popsáno asi 100 takových pravidel.

Čtvrtou částí popisu je seznam apelativních lemmat, která jsou přiřazena ke vzoru a případně i k typu alternace, pokud se daného apelativa alternace týká. Základ seznamu vznikl automatickou extrakcí apelativních lemmat ze slovníků staré češtiny (viz níže v části 2.2) a byl rozsáhle manuálně tříděn a upravován. K budování seznamu byl využit nástroj *OpenRefine* (2013, více viz níže v části 2.3). Seznam obsahuje asi 29 000 lemmat a ve spojení s ostatními částmi jej lze použít jako základ pro generování tvarů: ze seznamu lemmat získat tvarotvorné základy a na základě informace

10 Z praktických důvodů (zadávání z klávesnice) se ve formálním popisu pro nulový formant používá nula místo obvyklého znaku Ø.

11 Hláskové podoby lemmat vztažené k tomuto období uvádějí excerpované moderní slovníky staré češtiny.

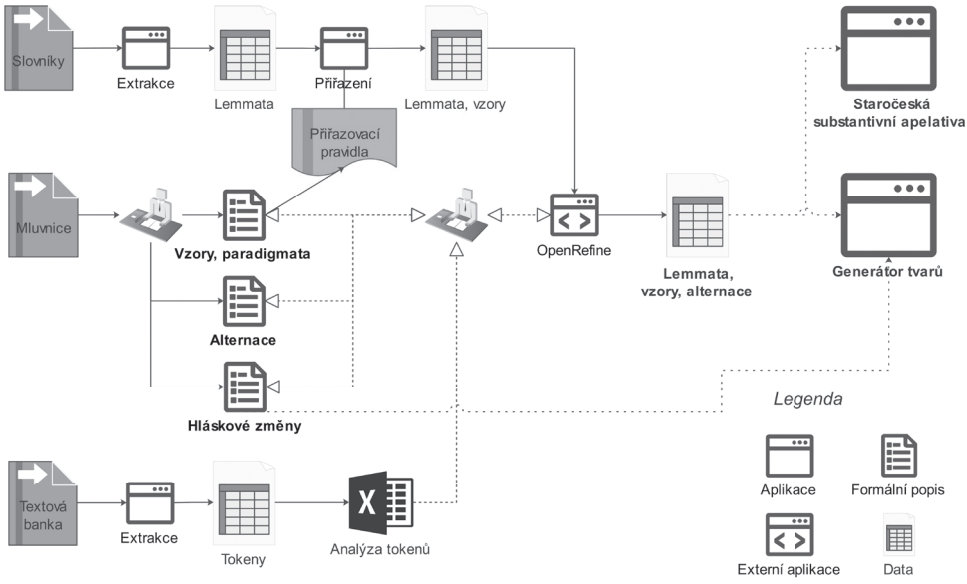
o vzoru je kombinovat se zakončeními (při zohlednění případných alternací). Pravidla pro formální změny zajistí formování tvarů podle fonotaktických a pravopisných pravidel i vytvoření všech pravidelných nástupnických podob.

Kromě těchto částí obsahuje popis seznam výjimečných tvarů (např. LOC.SG *časi* lemmatu *čas* nebo GEN.PL *číslóv* lemmatu *číslo*), jejichž systematické zavedení u všech zástupců daného vzoru by podklady zbytečně zatěžovalo.

První a druhá část popisu (deklinační typy a alternace), které se nejvíce týkají morfologie staročeského období, byly vydány knižně (Synková, 2017). Třetí a čtvrtá část, které jsou do značné míry technického rázu, jsou dostupné jako přílohy vydané práce na adrese <http://bit.ly/ridics-dm-book>. Jako příloha je zde dostupný také souhrn tabulek pro všechny vzory, seznam výjimek a seznam textů interní verze textové banky, na které proběhlo zkoumání většiny témat, s uvedením počtu tokenů a spolehlivosti/nespolehlivosti jednotlivých textů.

Výhodou zvoleného postupu při budování nástroje pro značkování (tagování) a lemmatizaci staročeských textů je vznik systematického popisu formální morfologie daného období a s tím související možnost využít v automatické morfologické analýze i detailní lingvistickou informaci (deklinační typ, hláskové změny). Nezbytnou cenou za tento přístup je časová náročnost popisu a jeho přímá závislost na zdrojích, s jejichž pomocí je budován. Zpracovaný popis tedy nutně představuje pouze základ, který bude s rozvojem použitých zdrojů třeba aktualizovat a dotvářet.

Na obecnější rovině práce testovala zvolený přístup (schematicky ho znázorňuje obrázek 1) jako celek — očekávalo se, že pokud na jejím základě vznikne úspěšný nástroj pro automatickou morfologickou analýzu staročeských apelativ, bude možné stejný/podobný postup použít i pro ostatní slovní druhy.



OBRÁZEK 1: Schéma zdrojů a aplikací využitých během přípravy formálního popisu staročeské apelativní deklinace a při vývoji následných aplikací.

3. ZPRACOVÁNÍ PODKLADŮ PRO FORMÁLNÍ POPIS MORFOLOGIE

3.1 *ANALÝZA TOKENŮ V EXCELU*

3.1.1 PŘEDSTAVENÍ NÁSTROJE A PŘÍKLADY VYUŽITÍ

K dohledávání staročeských tvarů zhruba 29 000 lemmat mezi několika miliony tokenů v textové bance byl využit nástroj *Analýza tokenů v Excelu* (2015), jehož hlavní předností je současná práce s množstvím tvarotvorných základů v různých morfonologických i pravopisných podobách (*hřiech-*, *hřieš-*, *hřích-*, *hříš-*; *otec-*, *otc-*, *votec-*, *votc-*, *ótc-*, *uotec-*…), které je nutno spojovat s netriviálním počtem zakončení včetně jejich nástupnických podob (*-u*, *-i*, *-iech*, *-ích*…). S využitím nulové koncovky lze pracovat rovněž s variantami slov jako *čso* a *co* i různými podobami nesklonných slov. Ve srovnání s korpusovým manažerem, jehož prostřednictvím se zpřístupňují data interní textové banky, nabízí tento nástroj úsporu času a zjednodušení práce s nelemmatizovanými daty (tvarotvorné základy jsou přehledně vyjmenované ve vstupní i výsledné tabulce; není nutné zadávat regulární výrazy). K jeho nevýhodám patří zejména to, že neuvádí slovo v kontextu, takže není možná desambiguace výrazu (tvar *ženu* je uveden pouze jednou, aniž se rozlišuje, zda se jedná o ACC.SG substantiva *žena*, nebo o 1.SG prézentu verba *hnáti*). Uvedený nástroj využívá volitelné součásti programu *Microsoft Excel*, jmenovitě *PowerPivot* (*Microsoft SQL Server 2012 SP2 PowerPivot pro Microsoft Excel* 2010, 2015; pro práci s velkými objemy dat v tabulkovém procesoru) a *Power Query* (*Microsoft Power Query pro Excel*, 2017; pro definici dotazů a manipulaci s daty v excelovském sešitu).[12]

Nástroj *Analýza tokenů v Excelu* pracuje se seznamem tokenů, z nichž každý je opatřen především informací o četnosti, zdrojovém textu a relevanci z hlediska bádání (viz níže) a údaji odvozenými přímo z tokenu. Každý zdrojový text je charakterizován zejména (pro účely nástroje) zkratkou literární památky (např. *BřezSnář* pro Snář Vavřince z Březové) a konkrétního pramene (*BřezSnářM* pro mikulovský rukopis, *BřezSnářS* pro stockholmský), datací pramene (okolo roku 1450, 1471 ap.), literárním druhem (próza, verš, drama), literárním žánrem a „spolehlivostí" jeho elektronické edice (zda už prošla finální redakční úpravou).

Údaj o relevanci tokenu se odvozuje z několika jeho vlastností. Za nerelevantní se považují tokeny v cizím jazyce, torzovitá, doplněná nebo rekonstruovaná slova, interpunkce a čísla.

Údaje odvozené z tokenu, které nástroj využívá při hledání i vytváření přehledů, jsou např. počet fonogramů (řetězců znaků, které odpovídají fonémům; jeden fonogram představuje např. *n*, *ch* a *ie*, takže výraz *nenie* obsahuje 4 fonogramy), jeden iniciální fonogram (*n*), jeden až tři[13] finální (*ie*, *nie* a *enie*) a podoba tokenu s obráceným pořadím fonogramů (*ienen*).

12 *PowerPivot* byl pro *Microsoft Excel 2010* k dispozici zdarma jako volitelný doplněk, od vyšších verzí se stal součástí programu, ale byl dostupný pouze v některých edicích, např. Professional.

13 Ukázalo se, že staročeské koncovky jsou tvořeny max. 3 fonogramy (např. podkoni*eho*).

Při popisu staročeské apelativní deklinace se využívaly především dvě funkce nástroje *Analýza tokenů v Excelu*. Za prvé se jeho pomocí hromadně vytvářely tvary a zjišťovala se jejich frekvence v textech interní textové banky, za druhé se v seznamech typů tvarů řazených podle posledních fonogramů hledaly doklady přejatých, netypických nebo jinde nedoložených koncovek u jednotlivých deklinačních typů.

Příkladem využití v první funkci je např. vytvoření a vyhledání všech tvarů staročeských *ъv*-kmenů. Jedná se asi o 70 lemmat, která mohou podle mluvnic v různé míře přijímat koncovky od *a*-kmenů, *ja*-kmenů i *i*-kmenů a zároveň mají pozůstatky svých původních tvarů. Aby bylo možné získat o deklinačním typu systematičtější přehled a ověřit mluvnické popisy, byly do nástroje zadány všechny tvarotvorné základy zástupců této deklinace (vybraných na základě historických mluvnic a staročeských slovníků) včetně všech formálních podob tvarotvorných základů (tedy např. *břěskev*, *břěskv*, *břeskev*, *břeskv* jako formální podoby tvarotvorného základu lemmatu *břěskev*) a všechny koncovky původní, *a*-kmenové, *ja*-kmenové a *i*-kmenové včetně svých nástupnických podob (na základě mluvnic a systémově i pro pády, ve kterých mluvnice dané koncovky neuvádějí). Nástroj z tvarotvorných základů a koncovek vytvořil tvary (zadané podmínky umožňovaly i vytvoření tvarů neočekávaných — např. **břeskeve*) a v dalším kroku zjistil, zda vůbec a s jakou frekvencí jsou doloženy v textech. Výsledek nabídl ve formě tabulky s lemmaty v řádcích a koncovkami ve sloupcích. Všechny nalezené tvary bylo potřeba zkontrolovat v textové bance kvůli případným homonymiím, doklady z nespolehlivých zdrojů poté i v předlohách. Vyhledat totéž přímo v textové bance by bylo nadmíru zatěžující a nadto nutně poznamenané možností chyby a nesystematičnosti. V tomto smyslu nástroj umožnil popis založit i na textové bance (a nespoléhat se pouze na mluvnice nebo malý soubor ručně excerpovaných textů).

Příkladem využití nástroje v druhé funkci může být např. hledání *i*-kmenových tvarů *ja*-kmenových lemmat, při kterém byly v seznamech typů tvarů řazených podle posledních fonogramů hledány např. mezi všemi tvary zakončenými na *-cech* tvary jako *ulicech*, *studnicech* (tj. *ja*-kmenová lemmata *ulicě*, *studnicě* s *i*-kmenovou koncovkou *-ech*), až došlo použitím různých filtrů a na jejich základě vyhledaných seznamů k pokrytí všech *i*-kmenových koncovek a všech možných zakončení tvarotvorného základu u *ja*-kmenů. I když byla míra ruční práce se seznamy značná, podobné hledání přímo v textové bance by znamenalo prohlédnout jednotlivé tokeny, ne jejich typy, čímž by náročnost práce vzrostla nad únosnou mez.

3.1.2 PŘÍPRAVA DAT

K vytvoření seznamu tokenů a pomocných seznamů/tabulek s informacemi (např. o jazyce tokenu, pravopisných invariantech ap.), s nimiž *Analýza tokenů v Excelu* pracuje, slouží extrakční program napsaný v jazyce C#. Jeho vstup tvoří jednak soubory s vertikálním textem (vertikálem[14]) jednotlivých pramenů, jednak samostatný seznam pramenů s názvem *Texty.txt* (o něm viz níže). Výstup extrakčního programu je vstupem pro *Analýzu tokenů*.

14 Srov. https://nlp.fi.muni.cz/cs/PopisVertikalu.

Požadovaný formát vertikálu je následující: Řádek každého tokenu se skládá ze sloupců oddělených tabulátorem. V prvním sloupci je uveden token, druhý až čtvrtý jsou vyhrazené pro hyperlemma, lemma a morfologickou charakteristiku, v pátém může být poznámka a v šestém jazyk. Rozpoznávané poznámky jsou (včetně hranatých závorek) *[torzovité slovo]*, *[doplněno]* a *[rekonstruováno]*. Na místě jazyka může být údaj *[cizí jazyk]*. Řádky se strukturními značkami jako *<doc>* (tj. začínající na <) se přeskakují.

Seznam pramenů (soubor *Texty.txt*) na každém řádku obsahuje ve sloupcích oddělených znakem | (svislou čárou) nejdříve pořadové číslo a název souboru (bez přípony). V dalších sloupcích následují název pramene a památky, slovní datace, z ní odvozené období vzniku (např. „1401–1450“), literární druh („próza“) a žánr („biblický text“) a nakonec spolehlivost („ano“, „ne“).

Data generovaná extrakčním programem tvoří několik textových souborů v obdobném formátu jako soubor *Texty.txt*. Primární informace o tokenech (jejich textová podoba, invariantní podoba, počáteční a koncové fonogramy) jsou uloženy v souboru *Tokeny.txt* a provázání tokenů s prameny (prostřednictvím pořadových čísel pramenů i tokenů) v souboru *Tokeny_v_textech.txt* (zároveň je tu uložena i frekvence tokenů). Ze souboru *Tokeny.txt* se kvůli úspoře místa odkazuje na odvozené informace v pomocných souborech opět prostřednictvím pořadových čísel; v souboru *Tokeny.txt* se tak např. místo hodnot „stará čeština“ a „cizí jazyk“ uvádějí čísla 1 a 2, přičemž samotné hodnoty jsou uloženy v souboru *Jazyky.txt*.

Další pomocné soubory vypadají obdobně: v souboru *Funkce.txt* jsou uvedeny funkce tokenu v textu (např. „interpunkce“ a „číslo“; tato informace mj. slouží k identifikaci nerelevantních tokenů), v souboru *Poznamky.txt* je textový obsah opakujících se poznámek s pořadovým číslem, které se uvádí v souboru *Tokeny.txt* místo hodnot jako „torzovité slovo“. Soubory *Invarianty.txt*, *Fonogramy1.txt*, *Fonogramy2.txt* a *Fonogramy3.txt* obsahují informace odvozené přímo z tokenů: tokeny s velkým počátečním písmenem,[15] fonogramy délky 1 (navíc s informací, zda jde o písmeno a je-li velké) a nakonec fonogramy délky 2 a 3.

3.2 PŘIŘAZENÍ VZORU K LEMMATU

Další nástroj, který byl naprogramován pro přípravu formálního morfologického popisu (s výhledem na pokrytí všech slovních druhů, nejen apelativ), sloužil k předběžnému přiřazení vzoru k lemmatu na základě údajů uvedených ve slovníku. Na vstupu byly údaje převzaté z moderních slovníků staré češtiny, tj. StčS, ESSČ, GbSlov a MSS, které byly k dispozici ve formátu XML. Ke každému lemmatu se shromáždily následující informace: heslové slovo, morfologická charakteristika, slovní druh, jmenný rod (u substantiv), vid (pro pozdější popis sloves), způsob zpracování,[16]

15 Tokeny rozlišují velikost písmen; pro jednotnou práci s podobou na počátku a uprostřed věty je nějaká normalizace nutná.

16 Údaj specifický pro ESSČ: zda bylo heslové slovo vyřazeno ze zpracování, neboť se jednalo o vlastní jméno nebo nenáležitou hláskovou podobu, která byla přeřazena k jinému heslu, nebo podobu doloženou po roce 1500 (tj. nestaročeskou) nebo heslo zpracované v MSS.

odkaz,[17] typ heslové stati,[18] zdroj (zkratka slovníku), číslo,[19] zda je lexikální jednotka nesklonná, význam, původní morfologická charakteristika, původní slovní druh. Výsledný seznam byl k dispozici ve formátu CSV s oddělovacím znakem tabulátorem.

Jelikož jednotlivé slovníky zpracovávají údaje o lemmatech odlišným způsobem, přistoupili jsme k následujícím úpravám údajů o heslových slovech: sjednotili jsme (pro snazší strojové zpracování) označení slovních druhů, resp. jmenných rodů a vidů (tj. např. zkratky „spoj." a „konj." jako „conj.", „ž." jako „f.", „dok." a „pftiv." jako „pf." ap.). Slovníky také uvádějí informace o heslových slovech v kompaktní formě, např. GbSlov zpracovává v jedné heslové stati (typograficky v jednom odstavci) výchozí heslo včetně zahnízdovaných odvozenin (obvykle adverbií u adjektiv: *krásně* v hesle *krásný*, ale i v jiných případech: *čujúcí* v hesle *čúti*), StčS po formálních změnách lexikografické koncepce v 80. letech (Němec, Nedvědová & Pečírková, 1980) zpracovává v odůvodněných případech v jedné heslové stati více lexikálních jednotek, např. perfektiva a imperfektiva (např. *popraviti* a *popravovati*), ESSČ zpracovává v jedné heslové stati heslovou podobu, která nabývala různých slovnědruhových funkcí (*aby* jako částice a spojka). Současné formální zpracování elektronické verze slovníků neumožňuje jednoznačné přiřazení uváděných údajů ke konkrétním podobám heslového slova, takže jsme při přípravě podkladů zvolili přístup, v němž byly víceznačné údaje o slovním druhu, morfologické charakteristice ap. nejprve sloučeny do jednoho údaje (jednotlivé části byly odděleny středníkem) a následně se pro každé heslové slovo jednotlivé údaje duplikovaly (např. lemma *artikul* s morfologickou charakteristikou „-e, pozd. též -a/-u" bylo uvedeno dvakrát: v podobě „artykul, -e" a „artikul, -a/-u").

Při výběru hesel ze slovníků do dalšího zpracování (tj. ze 153 314 heslových slov) jsme se snažili také o redukci údajů: nejprve jsme seznam omezili na substantiva s počátečním malým písmenem (tj. apelativa), ze seznamu byla rovněž vyřazena hesla z MSS bez morfologické charakteristiky, pokud bylo heslové slovo doloženo ve více slovnících, a dále jsme stanovili pořadí slovníků na základě kvality zpracování: StčS, ESSČ, GbSlov, MSS (pokud bylo např. lemma zachyceno v ESSČ a MSS, údaje z MSS se do dalšího zpracování nedostaly).

Posledním krokem při zpracování seznamu lemmat bylo přiřazení základního vzoru pomocí sady empiricky definovaných pravidel, která na základě údajů o rodu, zakončení v nominativu a zakončení v genitivu určila přiřazení k základnímu vzoru. Např. femininům končícím na *-ě* v nominativu i genitivu byl přiřazen základní vzor „subst.f.ja-kmen.duše". U málo zastoupených vzorů bylo jednodušší vyjmenovat konkrétní lemmata (tj. např. *dci* a *máti* pro vzor „subst.f.r-kmen.máti"). V případě mužských *o*-kmenů a *jo*-kmenů jsme do rozhodování o přiřazení ke vzoru zapojili klíčová slova významové definice (prvního významu heslové stati): pokud se např. u mas-

Tyto údaje slouží pro interní zpracování slovníku a kromě proprií se taková hesla nedostanou do publikované verze slovníku na stránkách *Vokabuláře webového*.

17 Odkazované heslo v případě, že šlo o odkazovou heslovou stať.

18 Zda se jedná o plnohodnotnou heslovou stať, odkazovou stať nebo stať, která není určena ke zveřejnění (pouze pro ESSČ).

19 Označení pluralií tantum.

kulin zakončených v nominativu na měkký konsonant a v genitivu na *-ě* v definici významu objevily výrazy jako *výrobce*, *řemeslník*, *osoba* ap., byl jim přiřazen vzor „subst.m.jo-kmen.muž“, pokud se však za podobných podmínek vyskytovaly ve významové definici výrazy *ryba* nebo *pták* ap., byl jako vzor přiřazen „subst.m.jo-kmen.sýc“. Pokud klasifikované lemma vyhovovalo několika definovaným pravidlům, byly k němu přiřazeny všechny vyhovující vzory (oddělené středníkem).

Výše uvedená pravidla byla definována ve strukturovaném dokumentu XML (viz obrázek 2).

```
<pattern name="subst.m.jo-kmen.sýc" gender="m">
    <sense-keywords filepath=".\subst.m.jo-kmen.sýc_sense.txt" />
    <rule nominativ="C" genitiv="ě" />
    <rule nominativ="l" genitiv="e" />
    <rule nominativ="řec" genitiv="rcě" />
</pattern>
<pattern name="subst.m.molle-adj.berúcí" gender="m">
    <rule nominativ="í" genitiv="ieho" />
</pattern>
```

OBRÁZEK 2: Ukázka pravidel pro přiřazení lemmatu ke vzorům („C“ slouží jako zástupný symbol pro konsonanty).

Takto zpracovaný seznam, který obsahoval 32 810 položek, prošel ruční kontrolou a přiřazením lemmat k jednotlivým specifickým vzorům. Např. pro apelativa základně patřící k ženským *ja*-kmenům (tedy s automaticky přiřazeným vzorem „subst.f.ja-kmen.duše“) bylo definováno 11 vzorů (jedním z nich byl i vzor duše, ale další vzory byly stanoveny pro jména typu *hospodyni*/*hospodyně*, *tvrz*/*tvrže*, jména jednoslabičná atd.). Duplicitní položky se z dalšího zpracování vypouštěly. Pro ruční kontrolu a přiřazování vzoru (a následně i značek pro alternace) byl využit program *OpenRefine* (viz dále).

Srovnání automatického přiřazení vzorů s ručně zkontrolovaným seznamem lemmat a vzorů ukázalo, že z celkového počtu 28 759 lemmat s přiřazeným vzorem bylo 13 941 položek shodných s jediným automaticky navrženým vzorem, v 3206 případech se finální vzor shodoval s jedním z několika automaticky navržených a v 11 612 případech byl konečný vzor odlišný od přiřazeného.

3.3 *OPENREFINE*

Aplikace *OpenRefine* (2013) byla již od samých začátků, kdy se o vývoj starala firma Google, určena pro analýzu a úpravu tabulkově uspořádaných dat. I když jsme při práci na formálním popisu využívali program *Microsoft Excel*, pro přiřazení lemmat ke vzorům jsme zvolili *OpenRefine* zejména z následujících důvodů: 1) pohodlná práce s velkým množstvím dat; 2) zaznamenávání jednotlivých (i hromadných) změn s možností návratu k předchozímu stavu; 3) výběr řádků pomocí filtrů a průřezů (v angličtině „facets“), které umožňují i vyhledávání pomocí regulárních výrazů;

4) různá označení jednotlivých řádků pro další zpracování (případně vyloučení řádku z dalšího zpracování bez nutnosti jej mazat); 5) programovací jazyk pro hromadnou změnu hodnot; 6) rychlé přehledy o počtu nalezených řádků při zvoleném zadání podmínek a výpisy jedinečných hodnot, které umožňují kontrolu jejich jednotného zadávání (např. kontrolu překlepů nebo jinak nejednotného zadání charakteristik).

Tento program funguje jako lokálně běžící webová aplikace, takže k editaci dat slouží webový prohlížeč. Sdílení dat je možné pomocí exportu do speciálního datového formátu programu. Údaje připravené v programu *OpenRefine* byly nakonec uloženy ve formátu XLS (*Microsoft Excel*), zejména pro lepší možnosti formátování, generování statistik, přehledů a pro transformaci do dalších formátů.

4. DATABÁZE STAROČESKÝCH APELATIVNÍCH VZORŮ

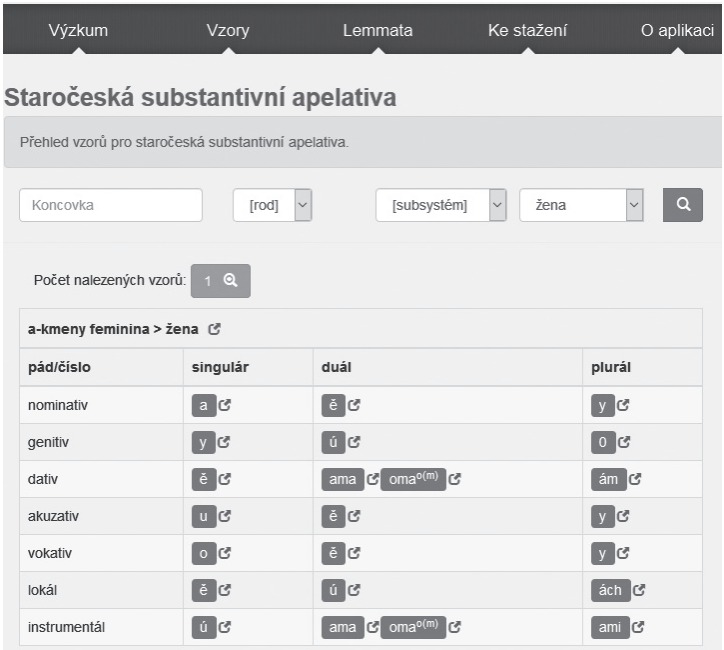
První aplikace, která se snažila využít a usouvztažnit údaje z formálního popisu staročeské apelativní deklinace, byla naprogramována s pomocí XML databáze *eXist-db* (2017). Tato webová aplikace nazvaná *Staročeská substantivní apelativa* (viz obrázek 3) umožňuje přistupovat k informacím směrem od seznamu vzorů k jednotlivým lemmatům i od lemmat k jejich vzorům a výskytům ve staročeské textové bance. V první části program zobrazuje paradigmata jednotlivých vzorů, přičemž jsou vždy barevně odlišena zakončení, která náležejí k substrátu, a zakončení, která jsou specifická pro daný vzor. U jednotlivých koncovek je také v indexu uveden jejich původ, pokud se jedná o koncovku jiného deklinačního typu, než ke kterému patří vzor (tj. např. o *o*-kmenovou koncovku u *a*-kmenového vzoru, viz koncovku *-oma* na obrázku 3). Paradigmata lze vyhledávat podle rodu, subsystému (tj. kmene, resp. indeclinabilií a výjimek), konkrétního vzoru nebo koncovky. Od jednotlivých vzorů lze přejít na stránku s lemmaty, která jsou ke vzoru přiřazena, a od jednotlivých zakončení lze přejít k jejich vyhledání ve staročeské textové bance.[20] Část věnovaná lemmatům rovněž umožňuje vyhledávat podle rodu, subsystému (tj. kmene, resp. indeclinabilií a výjimek) a konkrétního vzoru a dále podle podoby lemmatu (též s využitím regulárních výrazů). Od lemmatu lze jedním kliknutím přejít k jeho lexikografickému zpracování v moderních slovnících staré češtiny (ve *Vokabuláři webovém*),[21] popř. ke konkordanci nalezených tvarů lemmatu ve staročeské textové bance.[22]

Prvním krokem pro tvorbu uvedené aplikace byl převod vstupních dat do formátu XML. Vstupními daty byly 1) tabulky se substráty a ostatními zakončeními pro jednotlivé vzory (případně rovnou zakončení celých samostatných vzorů), 2) seznam

20 Prozatím se hledá pouze hlásková podoba zakončení k roku 1300, nikoli jejich nástupnické varianty. U frekventovaných zakončení (např. *-a*) může při zobrazení konkordančních řádků dojít vzhledem k velkému množství dat k chybě aplikace kvůli překročení časového limitu zpracování dotazu.

21 http://vokabular.ujc.cas.cz/hledani.aspx

22 Tato funkce není zatím zcela spolehlivá: neumí si poradit s alternacemi v tvarotvorném základu ani s formálními změnami tvarů během staročeského období.

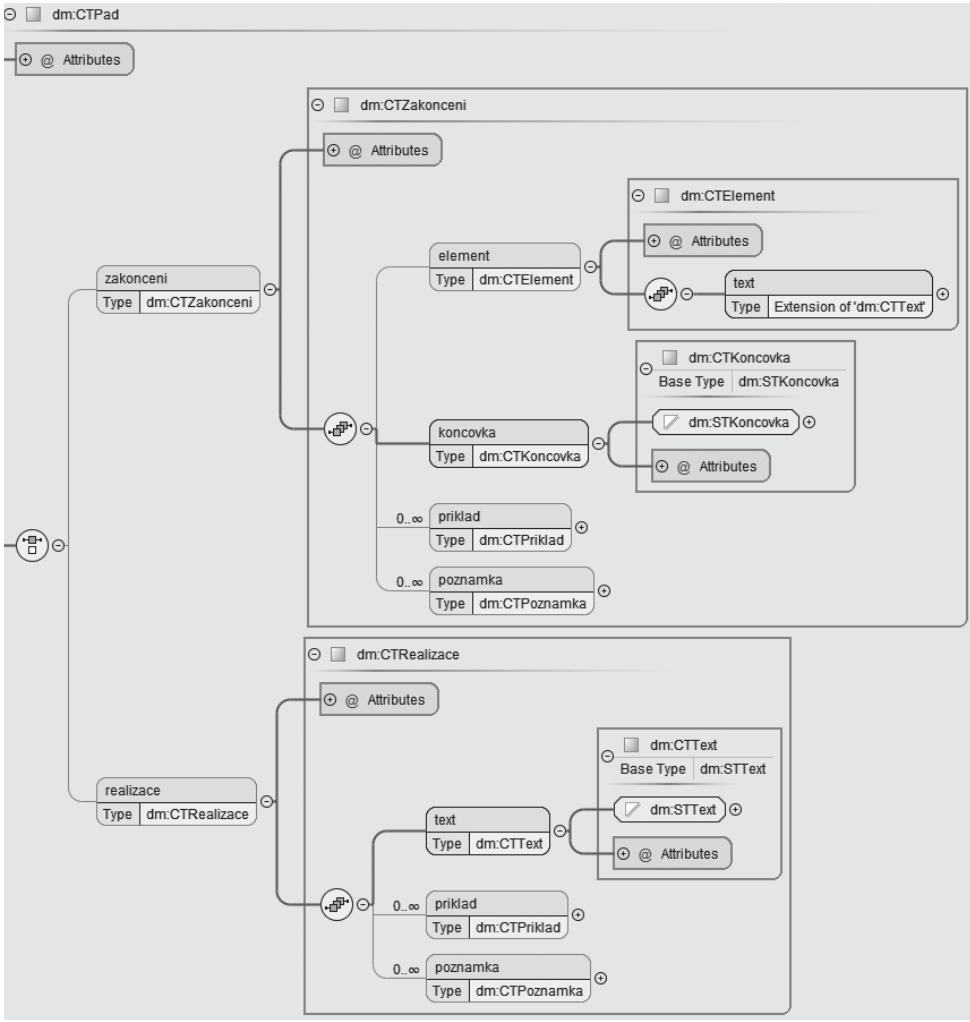


OBRÁZEK 3: Ukázka uživatelského rozhraní webové aplikace (stránka s formálním popisem vzoru *žena*).

apelativních lemmat (viz výše v části 2), tj. tabulka lemmat s přiřazeným vzorem a dalšími údaji: omezením čísla paradigmatu, případnou alternací v tvarotvorném základu, prvním významem lemmatu v heslové stati a zkratkou slovníku, z něhož bylo lemma převzato.

K převodu do požadovaného formátu sloužily XSLT transformace vstupních dokumentů. V prvním případě (zachycení paradigmat) byly relevantní části popisu označeny pomocí odstavcových a znakových stylů a pomocí samostatného nástroje byly z formátu DOCX převedeny do základního XML, které zachovalo tabulky a dvouúrovňové značkování (odstavce a dílčí úseky v rámci odstavců, tj. znakové styly). Na tento výstup se postupně aplikovalo několik XSLT transformací, až se podařilo docílit ideální struktury pro další zpracování. Výstup má hierarchickou strukturu: V rámci subsystému (např. *a*-kmeny, indeclinabilia) jsou definovány elementy označující rody (např. maskulinum), v jejich rámci pak substrát (repertoár zakončení společný pro více vzorů) a jeden nebo několik vzorů, případně samostatných vzorů a tzv. solitérů (tj. vzorů s jedním lemmatem). V rámci vzoru je definována deklinace, tvořená elementy pro jednotlivá čísla (singulár, duál, plurál), v jejich rámci je definováno sedm pádů. Jednotlivá zakončení mohou být tvořena několika prvky: koncovkou (např. NOM.SG kuř-*ě*), koncovkou s kmenotvorným elementem (např. GEN.SG kuř-*et*-*e*) a namísto zakončení může stát také element obsahující realizaci konkrétního tvaru, u něhož by se složitě definovala morfologická skladba (např. GEN.SG *téhodne* u vzoru *týden*), viz obrázek 4 a ukázka 1. Relevantní prvky formálního popisu

mají ve formátu XML identifikátor (uložený v atributu „id“) a dále odkazy na související elementy. Každý vzor např. obsahuje odkaz na korespondující substrát, označení rodu a subsystému, včetně jednoslovného označení názvu vzoru (tj. poslední složky jeho jedinečného identifikátoru). Tyto i některé další údaje (např. seznam různých zakončení, která se vyskytují v definici substrátu nebo vzoru) slouží k jednodušší tvorbě uživatelského rozhraní webové aplikace.



OBRÁZEK 4: Prvky, které se mohou vyskytnout při formálním popisu jednoho pádu, zachycené pomocí standardu XSD (*XML Schema Definition*).

```
- subsystem:
   _oznaceni: n-kmen
   rod:
   - _oznaceni: n
     _id: n.n-kmen
     substrat:
       _id: subst.n.n-kmen
       deklinace:
         singular:
           nominativ:
             zakonceni:
             - _rod: n
               koncovka:
                 _puvod: souhlaskova_deklinace
                 _id: subst.n.n-kmen.nom.sg.souhlaskova_deklinace.ě
                 _text: ě
             - _rod: n
               element:
                 _puvod: kmenotvorna_pripona
                 _id: nom.sg.kmenotvorna_pripona.en
                 _text: en
               koncovka:
                 _puvod: o-kmen_n
                 _id: subst.n.n-kmen.nom.sg.o-kmen_n.o
                 _text: o
           genitiv:
           ...
         dual:
           nominativ:
         ...
         plural:
         ...
     vzor:
     - _nazev: jmě
       _id: subst.n.n-kmen.jmě
       deklinace:
         singular:
           nominativ:
             zakonceni:
             - _rod: n
               element:
                 _puvod: kmenotvorna_pripona
                 _id: nom.sg.kmenotvorna_pripona.én
                 text:
                   _text: én
               koncovka:
                 _puvod: o-kmen_n
                 _id: subst.n.n-kmen.nom.sg.o-kmen_n.o
                 _text: o
```

UKÁZKA 1: Ukázka formálního popisu části substrátu a paradigmatu vzoru *jmě* (pomocí jazyka YAML[23]).

23 *YAML Ain't Markup Language*, viz http://www.yaml.org.

Tabulka s lemmaty a přiřazenými vzory se v programu *Microsoft Excel* uložila ve formátu „Tabulka XML 2003 (*.xml)" a na tento výstup se také postupně aplikovalo několik transformací popsaných v jazyce XSLT. Ve výsledku se u každého heslového slova uvádějí nejen údaje převzaté ze slovníků (zdroj a první výklad významu z heslové stati a formálně upravený popis alternace kmene), ale i údaje, které jej propojují s jednotlivými prvky formálního morfologického popisu (např. subsystém, rod a vzor).

Samotná webová aplikace využívá XML databázi *eXist-db*, její programové nástroje pro generování uživatelského rozhraní pomocí jazyka HTML a kaskádové styly Bootstrap[24] pro definici základního vzhledu aplikace. Ke generování výstupů ze dvou XML souborů (viz výše) slouží procedury napsané v jazyce XQuery.[25] Aplikace bude dostupná jako aplikační balíček[26] pro tuto XML databázi, díky čemuž lze zajistit její snadnou distribuci, instalaci i aktualizaci.

5. GENERÁTOR SLOVNÍCH TVARŮ, IMPLEMENTACE LEMMATIZÁTORU APELATIV

Primárním nástrojem, jenž vzniká na základě formálního popisu staročeské apelativní deklinace, je generátor slovních tvarů. Aplikace je naprogramována v jazyce C# s využitím platformy *.NET Core*.[27] Jejím úkolem je pro zadané lemma (např. *vajce*) vytvořit všechny podoby slovních tvarů, které existovaly nebo mohly existovat ve staročeském období (tedy DAT.PL *vajcóm*, později *vajcuom* a nakonec *vajcům* atd.). Vstupem je lemma, k němu přiřazený vzor (u lemmatu *vajce* „subst.n.jo-kmen.moře"), volitelné omezení paradigmatu (singulare nebo plurale tantum; u lemmatu *vajce* se neuplatňuje) a symbolický zápis morfonologického chování slova (u lemmatu *vajce* se vyskytuje značka alternace „zaklad-bezkoncovkove_tvary-vajec/tvary_s_koncovkou-vajc+vejc", která popisuje podobu tvarotvorného základu v GEN.PL s nulovou koncovkou (*vajec*) a tvarotvorné základy ve všech tvarech s nenulovou koncovkou (*vajc-* i *vejc-*)).

Prototyp nástroje dokázal postupně generovat tvary 96 staročeských vzorů (resp. zástupných lemmat) včetně všech jejich specifik: kromě omezení paradigmatu u vzorů jako *kamna* (kde je lemmatem tvar NOM.PL) bylo dále nutné pracovat nejen s koncovkami, ale i kmenotvornými sufixy — např. u vzoru *rámě* správně získávat tvarotvorný základ *ram-* i od lemmatu *ram-en-o*.

Souběžně s vývojem nástroje se vytvářela data pro tzv. regresní test. Data pro tento test obsahují pro každý vzor jednak seznam tvarů nějakého reprezentativního (dostatečně frekventovaného) lemmatu, jež jsou očekávány na výstupu generátoru (především tvary doložené ve staročeské textové bance), ale zároveň i takové tvary, které indikují chybu ve vstupních datech nebo v programu a jejich výskyt si vyžádá opravu. Např. před implementací formální změny (viz níže) *kě* > *cě* byl přidán očekávaný nesprávně utvořený tvar **hadačkě*. Podobně se postupovalo u některých dalších

24 https://getbootstrap.com/docs/3.3/css/
25 https://www.w3.org/XML/Query/
26 https://exist-db.org/exist/apps/doc/repo.xml
27 https://docs.microsoft.com/en-us/dotnet/core/

formálních změn, jejichž výsledkem je zrušení výchozího tvaru — jeden nebo několik vybraných tvarů se tím včas zařadilo mezi nesprávně utvořené (to platí např. u tvaru **veśma* lemmatu *ves*, kde je třeba počítat kvůli některým tvarům se značením měkkosti v tvarotvorném základu).

Už první lemmata (od *a*-kmenových vzorů) začleněná do testovacího souboru si vyžádala podporu pro alternace. Např. lemma *hadačka* se vyznačuje alternací Eø symbolicky zapsanou „Eø-bezkoncovkove_tvary-KeK/tvary_s_koncovkou-KK" (slovní popis značky viz v části 2). Za oba zástupné symboly K pro souhlásku se po odtržení nenulové koncovky *-a* dosadí *č* a *k*. Před vytvořením této části programu se u GEN.PL generoval nesprávný tvar **hadačk*, který se posléze stal součástí dat pro regresní test.

Dalším požadavkem na program byla podpora pro formální změny. Východiskem pro implementaci byla už existující programová knihovna[28] s několika pravidelnými změnami z historických mluvnic. Do ní jsou postupně začleňovány změny popsané v třetí části popisu staročeské apelativní deklinace (viz část 2). Původní verze knihovny a dat zahrnovala např. u změny *'u* > *i* kontext (po měkkých souhláskách), informaci (v současné době však nevyužívanou) o časovém rozpětí změny a příklady jednotlivých změn, pokud byly v historických mluvnicích k dispozici (mj. *zeḿu* > *zemi*).

Do modulu hláskových změn byla pro potřeby generátoru doplněna lokalizace změny (kdekoli uvnitř tvarotvorného základu, nebo jen na jeho konci; pouze v koncovkách ap.), negativní kontext (zákaz aplikace změny *'u* > *i* před *o*, tj. uvnitř diftongu *uo*, aby se netvořil tvar **koniov* od *koňuov*) a příznak zrušení tvaru, aby se do výstupu nedostalo např. **kózlě* (s koncovkou *-ě* od vzoru „subst.n.nt-kmen.kuřě"), ale až *kózle* (změna *lě* > *le* podle Gebauera (1963, s. 198) proběhla už před rokem 1300).

Jakmile program pokrýval všechny vzory, identifikoval a aplikoval většinu druhů alternací a ovládal nejnutnější formální změny, bylo možné začít s generováním tvarů z celého seznamu téměř 29 tisíc lemmat.

V této fázi vývoje už přestalo být efektivní ověřovat správnost generování tvarů (s dalšími alternacemi a hláskovými změnami) častým spouštěním testů, a proto bylo pro účel testování programu vytvořeno webové rozhraní, jež na vstupu přijímá jedno lemma s parametry (vzor, alternace, omezení paradigmatu) a vypisuje tabulku všech jeho tvarů (řazených do sloupců podle mluvnického čísla a do řádků podle pádu). Tvary, mezi nimiž došlo k formální změně, program v každé buňce tabulky zobrazuje uspořádané do stromu. U nástupnického tvaru se rovněž zobrazuje, jaká hlásková změna vedla k jeho vytvoření.

Protože se už při prvním spuštění generátoru s kompletním seznamem lemmat na vstupu nevyskytlo mnoho nedostatků (většina lemmat nepodléhá alternaci), byly vytvořeny aplikace, které už přímo přispívají k řešení úkolu počítačové lexikální analýzy staročeského textu (pro apelativní substantiva): nejdříve generátor morfologické databáze pro program *Majka* (Šmerk, 2009; Majka, 2017), jenž v ní rychle vyhledává k zadanému tokenu lemma a morfologickou značku (kódující slovní druh, jmenný rod, číslo a pád), a následně jednoduchý tagger, který v součinnosti s *Majkou* zpracuje vertikál korpusu a doplní ke známým tokenům (tj. možným tvarům apela-

28 Tj. samostatná součást počítačového programu, která poskytuje ucelenou sadu funkcí a lze ji využít i v jiných programech.

tiv) všechna možná lemmata a morfologické značky (bez pokusu o desambiguaci). K otestování validity a možností prohledávání takto anotované staročeské textové banky byl použit korpusový manažer *Manatee* (Rychlý, 2007; Manatee, 2017).

V současné době, kdy jsou pokryta všechna lemmata a druhy alternací, program generuje více než 3 miliony možných tvarů. Tento počet v blízké době ještě naroste, jakmile se doplní inventář formálních změn. Data regresního testu pokrývají 143 lemmat a 1078 různých slovních tvarů. Počet nesprávně utvořených tvarů, které reprezentují dříve se vyskytující chyby programu, je 46. Naopak prozatím negenerovaných tvarů, které zhruba odpovídají chybějícím hláskovým změnám, je 24 a jejich počet se stále zmenšuje.

6. SHRNUTÍ A VÝHLEDY

V článku byl představen formální popis staročeské apelativní deklinace, nové nebo existující softwarové nástroje použité při jeho přípravě a dvě aplikace, které na jeho základě dosud vznikly. Formální popis staročeské apelativní deklinace má čtyři části umožňující vytváření tvarů staročeských apelativ. Skládá se z repertoáru i) zakončení jednotlivých deklinačních typů (odpovídajících kmenům), ii) alternací, tj. změn tvarotvorných základů, které doprovázejí deklinaci a které nelze zavést formou pravidel přepisu určité sekvence písmen na jinou, protože jsou vázané na jednotlivá lemmata, iii) formálních změn, které je možné formou pravidla zavést (jedná se o změny hláskové i změny zohledňující fonotaktická a pravopisná pravidla), a iv) apelativních lemmat s přiřazeným vzorem a případně alternací, pokud se daného lemmatu alternace týká. Z těchto podkladů je možné odvodit tvarotvorné základy, na základě vzoru k nim při zohlednění případných alternací přidat zakončení a aplikací formálních změn získat jednak nástupnické tvary, jednak tvary zformované podle fonotaktických a pravopisných pravidel. Popis byl založen na historických mluvnicích, novodobých slovnících staré češtiny a tvarech vyskytujících se v textech interní textové banky.

Pro zjišťování frekvence tvarů v textové bance byl vytvořen nástroj *Analýza tokenů v Excelu* umožňující např. i) hromadné generování tvarů a hledání jejich výskytu v textové bance, ii) zobrazování tvarů ve smyslu typů řazených podle fonogramů na konci tvaru. Nástroj byl využit především při popisu zakončení deklinačních typů a při popisu alternací. Další nástroj byl vytvořen pro získání první verze seznamu lemmat. Umožnil přiřadit apelativní hesla z moderních staročeských slovníků k předběžnému vzoru. Ruční kontrola a rozřazení lemmat k definitivnímu vzoru proběhla v nástroji *OpenRefine*, který umožňuje práci s rozsáhlými seznamy.

Na základě formálního popisu vznikly dvě hlavní aplikace. První se jmenuje *Staročeská substantivní apelativa* a umožňuje zobrazovat jednotlivé vzory (i jejich soubory na základě rodu a subsystémů) a vyhledávat je podle koncovky, dále lze prohledávat lemmata a získat informaci o jejich vzoru a odkaz na lexikografické zpracování ve *Vokabuláři webovém* i na výskyty ve staročeské textové bance. Druhou aplikací je generátor tvarů zohledňující všechny informace zpracované ve formálním popisu staročeské apelativní deklinace. V současné době generátor pokrývá všechna lemmata, všechny alternace a značnou část formálních změn. Jeho výsledky již byly použity pro tagování a lemmatizaci textů staročeské textové banky.

Obě aplikace se během roku 2018 objeví na stránkách výzkumné infrastruktury RIDICS,[29] jejich zdrojová data i kód zveřejníme pod vhodnou licencí a otevřeme veřejnému vývoji na portálu GitHub pod účtem CzechLanguageInstitute.[30]

Další práce na lemmatizaci staročeských textů se zaměří na formální popis dalších slovních druhů (sloves, synsémantik, adverbií), začlenění výjimečných tvarů do formálního popisu v XML, zpřehlednění a vylepšení použité struktury XML a algoritmů pro generování náležitých tvarů ze vstupních dat, včetně zohlednění období vzniku textu při jeho tagování. Dále bude nutné najít způsob pro odhalení a eliminaci hláskově nenáležitých podob (např. s využitím regulárních výrazů pro detekci nemožných kombinací grafémů). Plánujeme rovněž vytvoření testovacího, ručně anotovaného korpusu staročeských textů, s jehož pomocí bude možné měřit úspěšnost morfologického značkování a lemmatizace.

LITERATURA

Bartoň, T., Cvrček, V., Čermák, F., Jelínek, T., & Petkevič, V. (2009). *Statistiky češtiny*. Praha: Nakladatelství Lidové noviny.

Černá, A., & Lehečka, B. (2015). *Metodika přípravy a zpracování elektronických edic starších českých textů*. Praha: Oddělení vývoje jazyka Ústavu pro jazyk český AV ČR, v. v. i. Dostupné z: <http://vokabular.ujc.cas.cz/moduly/nastroje/metodika/ke-stazeni>.

Daňhelka, J. (1957). Obecné zásady ediční a poučení o starém jazyce českém. In B. Havránek, J. Hrabák a kol. (Eds.), *Výbor z české literatury od počátků po dobu Husovu* (s. 25–35). Praha: Nakladatelství Československé akademie věd.

Daňhelka, J. (1963). Obecné zásady ediční a poučení o češtině 15. století. In B. Havránek, J. Hrabák, J. Daňhelka et al. (Eds.), *Výbor z české literatury doby husitské. Svazek 1* (s. 31–41). Praha: Nakladatelství Československé akademie věd.

Daňhelka, J. (1985). Směrnice pro vydávání starších českých textů. In *Husitský Tábor: sborník Muzea husitského revolučního hnutí Tábor*, *8*, 285–301.

ESSČ: *Elektronický slovník staré češtiny* [online] (2006–). Oddělení vývoje jazyka Ústavu pro jazyk český AV ČR, v. v. i. Dostupné z: <http://vokabular.ujc.cas.cz/hledani.aspx>.

GbSlov: Gebauer, J. (1970a). *Slovník staročeský. Díl I [A–J]*. Praha: Academia. Dostupné též z: <http://vokabular.ujc.cas.cz/hledani.aspx>.

GbSlov: Gebauer, J. (1970b). *Slovník staročeský. Díl II [K–N]*. Praha: Academia. Dostupné též z: <http://vokabular.ujc.cas.cz/hledani.aspx>.

Gebauer, J. (1960). *Historická mluvnice jazyka českého. Díl III. Tvarosloví. I. Skloňování*. Praha: Nakladatelství Československé akademie věd. Dostupné též z: <http://vokabular.ujc.cas.cz/moduly/literatura/>.

Gebauer, J. (1963). *Historická mluvnice jazyka českého. Díl I. Hláskosloví*. Praha: Nakladatelství Československé akademie věd. Dostupné též z: <http://vokabular.ujc.cas.cz/moduly/literatura/>.

Gebauer, J. (2007). *Historická mluvnice jazyka českého. Díl IV. Skladba*. Praha: Academia. Dostupné též z: <http://vokabular.ujc.cas.cz/moduly/literatura/>.

Komárek, M. (2012). *Dějiny českého jazyka*. Brno: Host.

Lamprecht, A., Šlosar, D., & Bauer, J. (1986). *Historická mluvnice češtiny*. Praha: Státní pedagogické nakladatelství.

29 http://vokabular.ujc.cas.cz/informace.aspx?t=ridics

30 https://github.com/CzechLanguageInstitute

MSS: Bělič, J., Kamiš, A., & Kučera, K. (1978). *Malý staročeský slovník*. Praha: Státní pedagogické nakladatelství. Dostupné též z: <http://vokabular.ujc.cas.cz/hledani.aspx>.

Němec, I., Nedvědová, M., & Pečírková, J. (1980). Problém rozsahu velkých historických slovníků a Staročeský slovník. *Slovo a slovesnost*, *42*(3), 238–248.

Rychlý, P. (2007). Manatee/Bonito — A Modular Corpus Manager. In P. Sojka & A. Horák (Eds.), *1st Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2007* (s. 65–70). Brno: Masaryk University.

Staročeská textová banka [online] (2008–). Verze dat 1.1.3 [citováno dne 1. 11. 2017]. Oddělení vývoje jazyka Ústavu pro jazyk český AV ČR, v. v. i. Dostupné z: <http://vokabular.ujc.cas.cz/banka.aspx>.

Staročeský slovník. Úvodní stati, soupis pramenů a zkratek (1968). Praha: Academia.

StčS: *Staročeský slovník. [Seš.] 1–26* (1968–2008). Praha: Academia. Dostupné též z: <http://vokabular.ujc.cas.cz/hledani.aspx>.

Synková, P. (2017). *Popis staročeské apelativní deklinace (se zřetelem k automatické morfologické analýze textů Staročeské textové banky)*. Praha: Filozofická fakulta Univerzity Karlovy.

Šmerk, P. (2009). Fast Morphological Analysis of Czech. In P. Sojka & A. Horák (Eds.), *Proceedings of Third Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2009* (s. 13–16). Brno: Masaryk University.

Trávníček, F. (1935). *Historická mluvnice československá*. Praha: Melantrich.

Vážný, V. (1964). *Historická mluvnice česká II. Tvarosloví. 1. část Skloňování*. Praha: Státní pedagogické nakladatelství.

Vokabulář webový [online] (2006–). Verze dat 1.1.3 [citováno dne 1. 11. 2017]. Oddělení vývoje jazyka Ústavu pro jazyk český AV ČR, v. v. i. Dostupné z: <http://vokabular.ujc.cas.cz>.

SOFTWARE

Analýza tokenů v Excelu (2015). Verze 1.0. Dostupné z: <http://vokabular.ujc.cas.cz/moduly/nastroje/analyza-tokenu/ ke-stazeni>.

OpenRefine (2013). Verze 2.6 beta 1. Dostupné z: <http://openrefine.org/download.html#openrefine-26-beta-1>.

Microsoft Excel 2010 (2010).

Microsoft SQL Server 2012 SP2 PowerPivot pro Microsoft Excel 2010 (2015). Verze 11.0.5635.3. Dostupné z: <https://www.microsoft.com/cs-cz/download/details.aspx?id=43348>.

Microsoft Power Query pro Excel (2017). Verze 2.49.4831.381. Dostupné z: <https://www.microsoft.com/cs-cz/download/details.aspx?id=39379>.

eXist-db (2017). Verze 3.5.0. Dostupné z: <https://bintray.com/existdb/releases/exist/3.5.0>.

Majka (2017). [bez verze]. Dostupné z: <https://nlp.fi.muni.cz/czech-morphology-analyser/>.

Manatee (2017). Verze 2.151.5. Dostupné z: <https://nlp.fi.muni.cz/trac/noske/wiki/Downloads>.

Pavlína Synková | Ústav českého jazyka a teorie komunikace FF UK
<pavlina.synkova@ff.cuni.cz>

Boris Lehečka | Ústav pro jazyk český AV ČR, v. v. i.
<boris@daliboris.cz>

Ondřej Svoboda | Ústav pro jazyk český AV ČR, v. v. i.
<svoboda@ujc.cas.cz>