ČASOPIS PRO MODERNÍ FILOLOGII 105, 2023, Č. 1, S. 121–140

Korpus DIA1900: jeho koncepce a vytváření[1]

Lucie Benešová – Karel Kučera – Kateřina Najbrtová –
Klára Pivoňková – Martin Stluka (Praha)

CORPUS DIA1900: ITS CONCEPTION AND BUILDING
The objective of the paper is to describe the principles for building the onemillionword DIA1900 Corpus consisting of Czech texts published between 1851 and 1900, designed to be both balanced and representative. There are two main goals determining the methods of corpus building and the decision to develop new tools tailored to the special needs of 19th century Czech: 1) to present the variability of Czech in the 2nd half of the 19th century (including spelling, morphology, wordformation) and 2) to link the detected variants to the appropriate lemmas. The paper presents the phases of the processing of the texts, including transcription, manual pre-annotation, as well as the construction of a large morphological dictionary and the selection of a suitable set of paradigms. Further sections are focused on annotation and morphological tagging and manual disambiguation. The objective was to create a gold standard, intended for use in the automatic annotation both of the DIA1900 corpus and the planned corpus of Czech texts of the years 1800–1850.

KEYWORDS
diachronic corpus, 19th century Czech, morphological dictionary, lemmatization, morphological annotation, tagset

KLÍČOVÁ SLOVA
diachronní korpus, čeština 19. století, morfologický slovník, lemmatizace, morfologické značkování, tagset

DOI
https://doi.org/10.14712/23366591.2023.1.8

1. ÚVOD

Tento článek má dva základní cíle. Prvním z nich je představení koncepce nově chystaného diachronního korpusu DIA1900, který je koncipován jako plně lemmatizovaný a morfologicky označkovaný korpus mapující češtinu 2. pol. 19. století a vzniká v Ústavu Českého národního korpusu (ÚČNK) FF UK. Druhým cílem je pak přiblížení procesů, které stojí v pozadí jeho vzniku.

Po stručném představení korpusu (oddíl 2) je samostatná část věnována zpracování textů (oddíl 3), odděleně je popsán proces předzpracování textového materiálu, tzv. předanotace (oddíl 4). Detailní pozornost je věnována morfologickému

1 Článek vznikl při realizaci projektu Český národní korpus (LM2018137) financovaného Ministerstvem školství, mládeže a tělovýchovy v rámci aktivity Projekty velkých infrastruktur pro VaVaI.

slovníku (oddíl 5) a jednotlivým fázím jeho vytváření, jako je rozřazování lemmat ke vzorům (oddíl 5.1) a generování paradigmat (oddíl 5.2), dále anotaci (oddíl 6) a desambiguaci (oddíl 7), v další části je prezentován samotný tagset pro morfologickou anotaci (oddíl 8). Korpus je inovativní v otázce zpracování proprií (oddíl 9). Článek končí shrnutím předností tohoto chystaného korpusu, v závěru jsou uvedeny i výhledy našich dalších cílů, kterých bychom rádi v návaznosti na tento korpus dosáhli (oddíl 10). Celý text tak umožňuje čtenáři komplexně nahlédnout do procesu budování nového diachronního korpusu.

2. KORPUS DIA1900

Jazykové korpusy představují nezpochybnitelný trend několika posledních desetiletí, přičemž co do počtu jednoznačně převažují korpusy synchronní nad diachronními.

Obecně lze k diachronním korpusům přistupovat dvěma způsoby. Lze budovat rozsáhlé korpusy (k nimž se svou koncepcí v rámci ÚČNK řadí i korpus DIAKORPV6[2]), založené na textech z několika staletí a obsahující miliony slov, ale bez náležité architektury a nástrojů, nebo je možné vytvářet korpusy rozsahem menší, zato komplexní a výběrem jazykových dat vyvážené.[3]

Tak je koncepčně tvořen i korpus DIA1900 obsahující 103 textů z 2. pol. 19. stol. (cca 1 mil. slovních tvarů). Jde o korpus lemmatizovaný a morfologicky značkovaný. Jednotlivé texty byly vybírány tak, aby byl vytvořen korpus reprezentativní a žánrově vyvážený, a to ve všech pěti desetiletích. Každá dekáda obsahuje vždy pokud možno rovnoměrně z hlediska počtu a rozsahu textů po třetině texty beletristické, žurnalistické a odborné. Lze očekávat, že žánrová a časová vyváženost poskytne v maximální možné míře komplexní obraz dobového jazykového systému. Výsledný korpus vystavěný tímto způsobem umožní kvalitní výzkum na různých jazykových rovinách.

Při vytváření korpusu DIA1900 bylo cílem co nejvěrněji zachytit a anotovat jazykovou realitu a variabilitu tak, aby byla co nejlépe a bez apriorních předpokladů připravena na lingvistické zkoumání, a zároveň ji zpřístupnit uživateli, který ji při vyhledávání nepředvídá.[4]

Dodržení tohoto cíle, nebo cílů dvou, byl přizpůsoben přístup ke zpracování materiálu — od zásad, podle nichž byly texty zpracovány, přes strategii při tvorbě roz-

2 Korpus je tvořen texty ze 14. až 20. století, jedná se o referenční, nereprezentativní korpus transkribovaných textů obsahující cca 4,1 mil. tokenů; není morfologicky anotovaný, tudíž umožňuje pouze vyhledávání prostřednictvím lexikálních jednotek či frází jako takových, což samo o sobě zásadně limituje komplexní jazykový výzkum.

3 Ideální typ, tj. rozsáhlý korpus zahrnující dostatečný počet textů ze všech období vývoje jazyka, lemmatizovaný a morfologicky značkovaný, je vzhledem k povaze materiálu aktuálně obtížně realizovatelný. Vyčerpávající charakteristiku diachronních korpusů různých jazyků a výstižné zhodnocení výhod a nevýhod jednotlivých přístupů nalezneme např. ve studii Petterssonové a Borina (2019).

4 Srov. Kučera et al. (2019).

sáhlého morfologického slovníku včetně modulu vlastních jmen až po usouvztažnění variant prostřednictvím našeptávače.[5]

3. ZPRACOVÁNÍ TEXTŮ

Samotná pravopisná soustava u textů z 2. poloviny 19. století neklade ve srovnání se staršími obdobími na zpracování příliš velké nároky. Zásadnější úkol představuje spíše posouzení a ošetření jazykových jevů, které jsou odlišné od současné podoby, resp. představují alternativu standardizované varianty.

Proto byla stanovena pravidla umožňující zachování informací o původní grafické podobě, buď za využití procesu emendace, nebo lemmatizace.[6] V korpusu DIA1900 zůstává v zásadě zachována podoba v prameni, rozdíl mezi transliterací a transkripcí[7] zpracovávaných textů je pro toto období minimální; tiskové chyby jsou označeny a zachovány v rámci emendačního režimu. Výjimku v tomto smyslu tvořilo několik oblastí, kde bylo v souvislosti s předpokladem pozdějšího automatického značkování rozhodnuto jinak (doplnění hranice slov v případech typu *-li, by (bylby > byl by)*, v oblasti velkých písmen a interpunkce).

Byly přitom využity značky, které sloužily k zachycení a uchování metajazykových informací týkajících se roku vydání, autora, titulu, názvu kapitoly a stránkování (<r>, <a>, <t>, <k>, <s>) i informací relevantních lingvisticky (<e>, <o>; detailně viz dále).

Kódy <o></o> byly využity pro označení jednoslovných i víceslovných cizojazyčných prvků, jež jsou pojímány jako citátové, tj. nejsou gramaticky zapojeny do okolního textu. V rámci lemmatizace pak každý token uvnitř tohoto označení získal specifický tag F---------- označující cizojazyčné citátové slovo bez určení dalších gramatických kategorií (detailněji viz dále).

Pro účely výchozí přípravy textů korpusu DIA1900 se jako jediná uspokojivá možnost ukázalo ruční zpracování (týmem zahrnujícím i externí vyškolené spolupracovníky)[8] — dobová rozkolísanost úzu a velká variabilita vyžaduje pečlivé posouzení

5 Nástroj, který nabízí uživateli při zadávání korpusového dotazu varianty synchronně nestandardní a potenciálně nepředvídatelné (typicky jde o jevy v oblasti pravopisu, alternace a vokalické kvantity, projevující se v celém paradigmatu).

6 Obecně by při přípravě diachronních korpusů (v raných fázích jazyka naléhavěji než v námi zpracovávaném období) z hlediska zachování grafické podoby pramene představovalo ideální řešení víceúrovňové zpracování zahrnující rovinu transliterovanou a transkribovanou. V době přípravy korpusu DIA1900 nebylo víceúrovňové zpracování technicky možné ani na této úrovni, ani na úrovni lemmatizace. Přesto se podařilo vypracovat koncept, který informaci o původní grafické podobě zachová a zároveň zpřístupní slovní tvary uživateli.

7 Jazykové zpracování textů pro korpus DIA1900 vychází ze standardního způsobu transkripce pro zpracování starších českých textů.

8 Tímto rozhodnutím se metoda zpracování korpusu DIA1900 odlišuje např. od postupu zpracování polského ručně anotovaného korpusu textů z období 1830–1918, který využil automatickou transkripci; srov. Kieraś — Woliński (2018).

z hlediska doložení daného jevu, a to nikoli jen na úrovni konkrétního slovního tvaru; žádá si vyhodnocení jevu v obecnějších vývojových souvislostech a verifikaci v relevantních slovníkových a korpusových zdrojích (dále SKZ).[9]

Z dnešního pohledu nestandardní slovní tvary byly tříděny na ty, 1) jejichž tvar byl pro účely lemmatizace upraven bez záznamu o provedené úpravě, 2) jejichž původní grafická podoba byla zachována v rámci emendačního kódu <e> a pro účely lemmatizace byl upraven slovní tvar, a na ty, 3) které byly zařazeny do procesu lemmatizace ve výchozí podobě.

(1) Bez záznamu o provedené změně bylo dnešním pravidlům přizpůsobeno psaní velkých písmen a dále některé jevy spojené s hranicí slov a interpunkce, zde ovšem s respektem k dobovému úzu a tendenci značit interpunkcí pauzy.

(2) U slovních tvarů, u nichž je způsob grafického záznamu zjevně chybný, dobově nedoložený, nejednoznačný nebo nezřetelný, byla grafická podoba v textu upravena a původní grafický záznam byl zachován v rámci emendačního kódu (např. *díváte <e> djwate</e>*, *těžké <e>težké</e>*, *uplynuly <e>uplinuly</e>*, *kráčel <e>krácel</e>*, *představuje <e>představnje</e>*, *veřejnému <e>veřejuému</e>*). Značka <e> tak byla využita pro původní podobu takového tvaru, který byl při zpracování upraven z jiného než formálně pravopisného hlediska; posun mezi původní a upravenou podobou (teoreticky) představuje fonologický/zvukový (např. *mě <e>me</e>*), gramatický (např. *tři měsíce uplynuly <e>uplinuli</e>*) či historicky podmíněný pravopisný rozdíl (např. psaní *y* po *c, s, z* po r. 1850 bylo emendačním kódem zaznamenáno jakožto již neadekvátní jev: *zasílateli <e>zasýlateli</e>*).

(3) V případech, kdy byla (grafická) podoba slovního tvaru, či jevu, který se v tomto tvaru realizoval, sice ze synchronního hlediska nestandardní, ale dohledatelná v dobových slovníkových a gramatických příručkách, případně byla v rámci korpusově zpracovaného i zpracovávaného textového materiálu dostatečně doložena (viz SKZ), zůstala tato podoba zachována bez emendace. V textu tak zůstaly neupraveny např. tvary typu *mathematika*, *saisona*,[10] *způsob*, *zákonník*, *příležitosť*, *ouřad*, *vokoun*, *vejsluní*.

9 Jako cenný zdroj údajů o jazykové variabilitě byly využívány dobové slovníky, především Jungmannův *Slovník česko-německý* (1835–1839) a *Česko-německý slovník* Františka Š. Kotta (1878–1893). Další zásadní zdroj nejen během zpracovávání textů, ale především ve fázi předanotace a obohacování morfologického slovníku (viz podrobněji dále) představovaly *Slovník spisovného jazyka českého (SSJČ)*, *Příruční slovník jazyka českého (PSJČ)* a *Akademický slovník cizích slov*. Za relevantní doklady byly považovány i jednotlivé excerpční lístky a doklady uvnitř výkladu heslového slova. Pro možnost rychlého přehledového vyhledávání byl využit slovníkový prohlížeč DEBDict a *Internetová jazyková příručka*; z korpusových zdrojů pak DIAKORPV6 a interní korpus 19. století.

10 U slov cizího původu lze sledovat širokou škálu pravopisné variability. Tato variabilita je daná nejen neustáleným užitím, ale odráží také postupné začleňování těchto slov do českého jazykového systému a lze předpokládat, že reflektuje také postup jejich fonetického přizpůsobení. Vzhledem k tomu, že je v této oblasti obtížné stanovit jednoznačné kritérium pro případnou úpravu pravopisného úzu, byly tvary v textu zachovány bez úprav a provázanost variant byla přesunuta na jiné nástroje využité při přípravě korpusu DIA1900.

Rozhodnutí zachovat v textu neupravené tvary konvenuje s cílem věrného zachycení dobových jazykových jevů. K tomu, aby byl zároveň dodržen cíl dohledatelnosti variant, resp. jejich provázanosti se synchronní či standardizovanou variantou, byly v návaznosti na typ variability vyvinuty specifické postupy s využitím ruční předanotace.

Slovní tvary odlišující se od standardizované varianty pravidelnými změnami vokalického charakteru (změna *ý>ej, ú>ou*, v omezenější míře také *é>í*) nebo protetickým *v* byly nagenerovány do morfologického slovníku v rámci paradigmatu standardizované varianty; srov. dále kap. 4. Tak byly zachovány a zároveň provázány se standardizovanou variantou prostřednictvím lemmatu, např. slovní tvar *vejsluní* získá interpretaci jako jeden z tvarů lemmatu *výsluní*, slovní tvar *vokoun* bude interpretován jako jeden z tvarů lemmatu *okoun* apod.[11]

V ostatních případech byly v rámci lemmatizace tvary standardní i tvary z dnešního pohledu nestandardní nagenerovány do morfologického slovníku samostatně (viz dále).[12]

4. PŘEDANOTACE

Ke slovním tvarům odlišujícím se od standardu byly v procesu předanotace připojeny informace o lemmatu (lemmatech) a případně také morfologická charakteristika ve formě standardního Hajičova tagu (Hajič, 2004) ve zkrácené podobě; zpravidla bylo využito několik prvních pozic.[13] Pokud bylo možné daný tvar připojit k více lemmatům, byla všechna taková lemmata zaznamenána (jako první byla vždy uváděna podoba lemmatu odpovídající konkrétnímu tvaru).[14]

11 Toto řešení bývá tradičně označováno jako hyperlemma https://wiki.korpus.cz/doku.php/pojmy:lemma?s[]=hyperlemma; např. tvary *bejt* a *být* mají v celých svých paradigmatech přirazeno lemma *být*. Nejde zde přitom o koncepci víceúrovňové lemmatizace, jak byla představena u nejnovějších korpusů řady SYN.

12 Pro ruční desambiguaci je počítáno s propojením jevů homonymních a gramaticky a kontextově/významově zcela zaměnitelných. Tam, kde u tvaru není možné ani z kontextu rozhodnout, kterou z možných alternativ lemma + tag zvolit, budou uvedeny obě rovnocenné interpretace (ve spojení *podivil se nemálo jeho pracovitosti* získá výraz *pracovitosti* při desambiguaci interpretaci s vícečlenným lemmatem *pracovitost/pracovitosť N-FS3-----A-----*; srov. též poznámku 29). Navíc byly texty ošetřeny ruční předanotací, která umožnila, aby samostatně nagenerovaná lemmata byla se standardizovanými variantami propojena mimo morfologický slovník prostřednictvím dodatečného nástroje — našeptávače, který bude uživateli na úrovni lemmatu signalizovat provázanost typu *mathematika — matematika, saisona — sezóna, způsob — způsob, zákonník — zákoník, příležitosť — příležitost*.

13 Naše morfologické značení v rámci předanotace bylo „pouze" pracovní a návazné na užitý tagset (viz dále).

14 Ukázky různých typů značení tvarů odlišných od standardu: a) substantiva českého a jinojazyčného původu (*„„"vplyvu--vplyv,vliv;N"""*, *„„"charlatanerii--charlatanerie,šarlatanérie;N"""*), b) adjektiva (*„„"reálních--reální,reálný;A"""*, *„„"gummová--gummový,gumový;A"""*), c) tvary sloves (*„„"způsobila--způsobiti,způsobiti;Vp"""*, *„„"odpovědíti--odpovědíti,odpověděti;*

V těch případech, kdy se slovní tvar odlišoval od standardu[15] pouze pravidelnou hláskovou změnou (viz výše), k němu bylo připojeno standardizované lemma a slovnědruhová charakteristika, a to opět ve formě prvních pozic Hajičova tagu.[16]

Předanotace probíhala v textovém editoru Atom,[17] který se vyznačuje velkou mírou flexibility nastavení. Jejím účelem bylo vyrovnat se s nestandardními pravopisnými, hláskovými a morfologickými variantami charakteristickými pro jazyk 19. století a zachytit je s co nejmenším zkreslením a zároveň v takové podobě, aby byly pro uživatele korpusu snadno odhalitelné, popř. pomocí standardizovaných podob v budoucnu dohledatelné. Detailní charakteristice nestandardních variant jsme věnovali článek Kučera et al. (2019).

Vlastní předanotace byla nejprve realizována s důrazem a ohledem na novočeskou podobu problematických jevů. Účelem této předanotační fáze bylo zachytit jevy bezpečně dobově doložené — stala se jedním ze zdrojů tvarů a lemmat pro morfologický slovník.

Tímto způsobem byly zachyceny jednak odlišnosti v konkrétní lexikální jednotce jako takové, jednak odlišnosti v konkrétních tvarech lexikálních jednotek (nestandardní kmenotvorné přípony, koncovky s provedenou hláskovou změnou apod.).

V první fázi byla označena např. atypická vlastní jména, cizí slova (či morfémy) nedoložená v SKZ nebo slova vytvořená z dnešního pohledu nestandardními slovotvornými procesy. Podobně byla anotována slova obsahující některou z hláskových změn (*ý>ej; é>í; ú>ou;* protetické *v*; zjednodušení skupiny *konsonant + l*; měkčení v participiích a imperativech), případně z pohledu synchronního jazyka již nestandardní částici (*-ť, -tě, -ž, -že*) nebo koncovku odlišného paradigmatu; dále pak slova s neobvyklou, ale systémovou alternací. Jiným typem značení byly identifikovány jednoznačné odchylky v kvantitě či pravopisně chybné shody podmětu s přísudkem a nesystémové hláskové/pravopisné odchylky v rámci konkrétního slovního tvaru.

Modifikovanou morfologickou specifikaci si vyžadovaly také případy týkající se nestandardní hranice slov, kdy druhý člen dvojice představoval slovnědruhově neurčitou část budoucí spřežky. Bez zvláštního doplňujícího značení byly ponechány jevy,

Vf“““), d) adverbia (*„““poznenáhla--poznenáhla,poznenáhlu;D“““*, *„““vezpod--vezpod,vespod;D“““*).

15 Za nestandardní jazykové jevy byly považovány všechny výskyty, které označil spell checker. Kvůli značnému počtu nestandardních jevů různého typu a jejich celkové nepředvídatelnosti byl proces předanotace prováděn ručně, nikoli automaticky. K obdobnému závěru dospívají také Linde a Mittmann (2013, s. 244): ruční předanotace zajišťuje vysokou spolehlivost dat, přičemž v tomto kroku nelze prozatím počítat s předanotací automatickou. (Korpusy, u nichž byla využita automatická předanotace, vyžadují důkladnější kontrolu než korpusy, u kterých žádná (sic!) předanotace neproběhla.)

16 Ukázky značení nejběžněji se vyskytujících tvarů lišících se od standardu pravidelnou hláskovou změnou: a) *ý>ej* (*„““bejkovcem--H:býkovec;N“““*), b) *ú>ou* (*„““outraty--H:útrata;N“““*, *„““bezouhonné--H:bezúhonný;A“““*), c) protetické *v* (*„““vočkoval--H:očkovati;Vp“““*), d) *é>í* (*„““zelé--H:zelí;N“““*).

17 *Atom. A hackable text editor for the 21st Century* [online]. [cit. 29. 3. 2022]. Dostupné z: <https://atom.io/>.

které se objevují v rámci velkého množství paradigmat (*-ové* v Npl mask., *-ův* v Gpl mask., *-iž, -ž* v imperativu, *-nul* u vzoru *tisknout*; *-ě(e)jí* u vzorů *prosí* a *trpí*); tyto byly implementovány přímo do jednotlivých vzorů a jako jejich součást se staly obsahem morfologického slovníku.

5. MORFOLOGICKÝ SLOVNÍK

Cílem vytvářeného morfologického slovníku (dále MS) je pokrýt rozsáhlou jazykovou variabilitu[18] češtiny 19. století tak, aby slovník bylo možné použít jako spolehlivý zdroj pro morfologické značkování zdrojových textů a následnou ruční desambiguaci etalonu. Výsledný etalon (tj. manuálně morfologicky anotovaný pětisetisícový korpus) poslouží jako základ pro tagger a automatickou morfologickou anotaci.

MS obsahuje dva základní zdroje dat. Jedním z nich jsou současné i dobové lexikografické a v menší míře i mluvnické publikace (viz SKZ). U SSJČ a PSJČ bylo možné automaticky využít celé hesláře těchto slovníků včetně morfologických charakteristik. Na základě převzatých slovníkových údajů o morfologické charakteristice (u substantiv jmenný rod a koncovka genitivu singuláru, u některých hesel také nominativu plurálu, u sloves pak vid) byla provedena základní poloautomatická analýza. Například seznamy extrahovaných substantivních lemmat byly rozřazeny podle jednotlivých rodů, tj. na maskulina životná a neživotná, neutra a feminina. Následně byl tento seznam retrográdně uspořádán, což pomohlo při ručním přiřazování jednotlivých lemmat ke vzorům.[19] Data tohoto základního slovníku pak byla dále postupně rozšiřována, každé doplněné lemma bylo uvedeným procesem přiřazeno ke vzoru. Ačkoli jsou slovníky jako zdroje dat relativně bohaté, jedná se i tak o do značné míry nekompletní informace: neobsahují kompletní paradigmata nutná k morfologické analýze a nejsou v nich standardně uváděny některé typy lemmat, jako např. verbální substantiva, dějová a posesivní adjektiva nebo adjektiva tvořená od zeměpisných jmen příponami *-cký* a *-ský*.

Druhý podstatný zdroj tvořil textový materiál, který byl v rámci předanotace zpracováván tak, aby byly podchyceny jevy jazykově variantní a mohly pak být včle-

18 Vzhledem k tomu, že v době vytváření koncepce lemmatizace pro korpus DIA1900 ještě dvojúrovňová lemmatizace tak, jak je řešena např. v korpusu SYN2020, nepřicházela v úvahu, nebylo jí možné na zachycení variability použít. Přesto se podařilo vypracovat koncepci, která i při jednoúrovňové lemmatizaci představuje plastický model usouvztažnění variant, který navíc zohledňuje typologii jazykových jevů, které se ve variantách odrážejí. K tomuto účelu využíváme přidělení synchronně standardního lemmatu, resp. hyperlemmatu, dále možnosti informovat uživatele o existenci variant prostřednictvím našeptávače, a také kategorii vícečlenného lemmatu (tj. způsob zachycení tvarové homonymie u významově rovnocenných variant, kde ani z kontextu není možné o lemmatu rozhodnout; např. tvaru *brambory* je přiděleno současně lemma a tag *brambora N-FP4-----A-----* i *brambor N-IP4-----A-----*).

19 I ve fázi manuálního rozřazování k jednotlivým vzorům jsme přihlíželi k sémantické i morfologické charakteristice těchto slovníkových lemmat.

něny do MS. Tyto jazykově variantní jevy mají ve výsledném slovníku buď kompletní vlastní paradigma,[20] případně byl do již existujícího (nagenerovaného) paradigmatu doplněn jen některý z nestandardních tvarů. Výsledný formálně lexikální základ slovníku je tedy v porovnání s tradičními popisy v dobových gramatikách a slovnících v mnoha ohledech detailnější, jelikož je nutné, aby kompletně postihl pokud možno všechny lexikálně-gramatické struktury, které nejsou při popisu v uvedených publikacích systematicky zohledňovány. Základ pro generování MS byl tvořen lemmatem zařazeným k adekvátnímu vzoru a v rámci něj připraveným k nagenerování.

5.1 VZORY

Pro vytváření MS bylo nutné zvolit adekvátní množství vzorů,[21] které by beze zbytku umožňovaly formální popis jazyka 19. století v komplexní jazykové variabilitě. Základní systém vzorů byl vytvářen poměrně dlouho před samotným započetím práce s daty, konečný počet vzorů i jejich podoba se následně formovaly a doplňovaly i během ručního rozřazování. Jednotlivé vzory[22] se snaží reflektovat všechny typy morfologických alternací u substantiv všech rodů, např. krácení a dloužení vokálů (*kráva>krav*, *jméno>jmen* apod.), vkládání nebo naopak vypouštění vokálů (*malba>maleb*, *chrchel>chrchle*, *lev>lva*, *jídlo>jídel*, *pero>per* apod.), alternace kořenné samohlásky v nepřímých pádech (*ů>o*, např. *důl>dolu*, *bůh> boha*, *hnůj>hnoje*, *hůl>holi*, *í>ě dílo>děl* apod.), alternace vzniklé v důsledku palatalizace velár (*ch>š břicho>břiše*, *g>z kolega->kolezích*, *k>c páka>páce* apod.) a samozřejmě i různorodé kombinace všech výše uvedených alternací (*lebka/lebce* — alternace *k>c* + vkladné *e* v Gpl *lebek*), *spánek/spánku/spáncích*, *středisko/střediskách/střediscích* (alternace *k>c* + varianty *-e-/-o* v Gpl) apod.); to vše v různých deklinačních typech. U adjektiv je tomu obdobně (četné alternace v Npl, např. *ký>cí měkký*, *hý>zí nebohý*, *chý>ší hluchý*, *rý>ří starý*, *ský>ští ruský*, *cký>čtí laický*). Samostatné vzory mají i přejatá slova se specifickým skloňováním, např. mas-

20 Vlastní paradigma bude mít takový lexém, u nějž se variantnost (nestandardnost) projevuje v rámci celého paradigmatu. K tomu srov. Hlaváčová et al. (2019), kde je tento jev popsán v rámci pojmů flektivní a paradigmatická variabilita; dále srov. Osolsobě et al. (2017), kde jsou v této souvislosti zmiňovány globální a flektivní mutace.

21 Týká se substantiv, adjektiv a verb. Pro generování proprií, která mají v rámci korpusu DIA1900 vlastní tag, byly u substantiv i adjektiv využity vzory apelativní; tam, kde to možné nebylo, byly vytvořeny vzory nové.

22 Vytvořené vzory nemají za cíl primárně reflektovat systém substantivní nebo jiné jmenné flexe v češtině (ať už současné nebo v některých odlišnostech existující v průběhu 19. století). Jsme si vědomi toho, že k řadě procesů, jako je např. alternace kořenných hlásek, nedochází v české substantivní flexi nahodile a že je možné tyto pravidelnosti zohlednit při vytváření vzorů, což se dělo. Nicméně naše pojetí vzorů můžeme nazvat kombinací morfologicko-lexikálního přístupu (to je také důvod, proč máme poměrně velké množství vzorů). Podstatou našeho přístupu je požadavek, aby se podle daného vzoru dala bez dodatečných pravidel a zásahů nagenerovat všechna slova vykazující různou kombinaci alternací v rámci paradigmatu (podle našeho vzoru *lebka* je nagenerováno např. slovo *trubka* (alternace v Lsg a vkladné *e* v Gpl), nikoli však např. slovo *keramika*, které má sice tutéž alternaci v Lsg, ale v Gpl je bez vkladného *e*); na nagenerování slova *keramika* máme vzor *páka*.

kulina typu *cyklus*, *radius*, *genius*, *dingo*, feminina typu *nausea*, *rachitis*, *synopsis*, *tibia*, neutra typu *centrum*, *epiteton*.

Stejně je tomu i v případě sloves. Systém slovesných konjugací je v češtině velmi složitý. Velký počet výsledných slovesných vzorů byl nutný do značné míry též proto, že při automatickém generování jednotlivých tvarů je i u formálně blízkých sloves třeba počítat s různými hláskovými změnami v systémově totožně tvořených tvarech (např. *vystřihnout*: *vystřihnut/vystři**že**n*, ale *vytisknout*: *vytisknut/vyti**št**ěn*).

Všechny seznamy vzorů pak byly dále doplněny o nepravidelné tvary, v případě substantiv a adjektiv také o tvary nesklonné. Počet vzorů je díky tomu poměrně vysoký (detailněji viz tabulka 1 a 2).

Počet **substantivních** vzorů (bez variantních modifikací pro propria)	**164**
Počet **slovesných** vzorů	**294**
Počet **adjektivních** vzorů (bez variantních modifikací pro propria)	**24**

TABULKA 1. Počty jednotlivých vzorů

	N-M	**N-I**	**N-F**	**N-N**
Počet substantivních vzorů	49	38	49	28

TABULKA 2. Počty jednotlivých vzorů podle rodů[23]

5.2 GENEROVÁNÍ PARADIGMAT

Abychom mohli dokončený slovník použít pro samotnou morfologickou anotaci, bylo nutné zdrojový textový materiál slovníku roztřídit k vytvořeným vzorům (u ohebných slovních druhů; viz výše) a poté ho rozgenerovat do kompletních paradigmat pomocí jednoduchých scriptů[24] vytvořených v rámci diachronní sekce ÚČNK. Scripty byly koncipovány tak, aby obsahovaly typické paradigma pro daný vzor, v němž bylo ke každému tvaru přiřazeno lemma a potřebná morfologická informace, resp. veškerá možná morfologická interpretace daného tvaru (ve formě šestnáctimístného morfologického tagu).

Ukázka scriptu ke generování vzoru **žák**

```
writeln(fout,(lemma+#09+hyper+' '+'N-MS1-----A-----'));
writeln(fout,(kmen1+'a'+#09+hyper+' '+'N-MS2-----A-----'+' '+'N-MS4-----A-----'));
writeln(fout,(kmen1+'u'+#09+hyper+' '+'N-MS3-----A-----'+' '+'N-MS5-----A-----'+' '+'N-MS6-----A-----'));
writeln(fout,(kmen1+'ovi'+#09+hyper+' '+'N-MS3-----A-----'+' '+'N-MS6-----A-----'));
writeln(fout,(kmen1+'em'+#09+hyper+' '+'N-MS7-----A-----'));
writeln(fout,(kmen2+'i'+#09+hyper+' '+'N-MP1-----A-----'+' '+'N-MP5-----A-----'));
writeln(fout,(kmen1+'ové'+#09+hyper+' '+'N-MP1-----A-----'+' '+'N-MP5-----A-----'));
```

23 N-M: životná maskulina, N-I: neživotná maskulina, N-F: feminina, N-N: neutra.

24 Každý ze vzorů byl zpracován jedním scriptem.

```
writeln(fout,(kmen1+'ů'+#09+hyper+' '+'N-MP2-----A-----'));
writeln(fout,(kmen1+'ův'+#09+hyper+' '+'N-MP2-----A-----'));
writeln(fout,(kmen1+'ům'+#09+hyper+' '+'N-MP3-----A-----'));
writeln(fout,(kmen1+'y'+#09+hyper+' '+'N-MP4-----A-----'+' '+'N-MP7-----A-----'));
writeln(fout,(kmen2+'ích'+#09+hyper+' '+'N-MP6-----A-----'));
writeln(fout,(kmen1+'ama'+#09+hyper+' '+'N-MP7-1---A-----'));
```

Ukázka nagenerované části jednoho slovesného paradigmatu (lemma **odmetat**, vzor **dělat**) s šestnáctimístným morfologickým tagem

```
neodmetal       odmetat VpIS------NA---I
neodmetal       odmetat VpMS------NA---I
neodmetala      odmetat VpFS------NA---I
neodmetala      odmetat VpNP------NA---I
neodmetalas     odmetat_být VpFS------NA-1-I_VB-S---2--AA-1-I
neodmetalatě    odmetat VpFS------NA--TI
neodmetalatě    odmetat VpNP------NA--TI
neodmetalať     odmetat VpFS------NA--TI
neodmetalať     odmetat VpNP------NA--TI
neodmetalaťs    odmetat_být VpFS------NA-1TI_VB-S---2--AA-1-I
neodmetali      odmetat VpMP------NA---I
neodmetalitě    odmetat VpMP------NA--TI
neodmetaliť     odmetat VpMP------NA--TI
neodmetalo      odmetat VpNS------NA---I
neodmetalos     odmetat_být VpNS------NA-1-I_VB-S---2--AA-1-I
neodmetalotě    odmetat VpNS------NA--TI
neodmetaloť     odmetat VpNS------NA--TI
neodmetaloťs    odmetat_být VpNS------NA-1TI_VB-S---2--AA-1-I
neodmetals      odmetat_být VpIS------NA-1-I_VB-S---2--AA-1-I
neodmetals      odmetat_být VpMS------NA-1-I_VB-S---2--AA-1-I
neodmetaltě     odmetat VpIS------NA--TI
neodmetaltě     odmetat VpMS------NA--TI
neodmetalť      odmetat VpIS------NA--TI
neodmetalť      odmetat VpMS------NA--TI
neodmetalťs     odmetat_být VpIS------NA-1TI_VB-S---2--AA-1-I
neodmetalťs     odmetat_být VpMS------NA-1TI_VB-S---2--AA-1-I
neodmetaly      odmetat VpFP------NA---I
neodmetaly      odmetat VpIP------NA---I
neodmetaly      odmetat VpNP------NA---I
neodmetalytě    odmetat VpFP------NA--TI
neodmetalytě    odmetat VpIP------NA--TI
neodmetalytě    odmetat VpNP------NA--TI
neodmetalyť     odmetat VpFP------NA--TI
neodmetalyť     odmetat VpIP------NA--TI
neodmetalyť     odmetat VpNP------NA--TI
```

```
neodmetat       odmetat Vf--------NA---I
neodmetati      odmetat Vf--------NA---I
neodmetatis     odmetat_být Vf--------NA-1-I_VB-S---2--AA-1-I
neodmetats      odmetat_být Vf--------NA-1-I_VB-S---2--AA-1-I
neodmetav       odmetat VmFP------NA---I
```

Další krok představovala revize nagenerovaných paradigmat v rámci zvolených vzorů. Poloautomatickým způsobem byly identifikovány a revidovány chybně nagenerované položky, které vznikly například neadekvátním zařazením daného lexému k určitému vzoru nebo formální chybou při tvorbě scriptu.[25] Proces revize zahrnoval také doplnění paradigmat u některých lexikálních jednotek.[26]

6. ANOTACE

Za pomoci nagenerovaného MS byla provedena anotace zdrojových dat,[27] přičemž každému tokenu byla přidělena náležitá interpretace obsahující dvojici lemma a tag. Systém přiřazování byl nastaven tak, aby bylo ke každému tokenu přiřazeno tolik interpretací, kolik jich daný token může v jazykovém systému mít. To ve výsledku znamená, že nehomonymní tokeny získají pouze jednu náležitou interpretaci; oproti tomu tokenům homonymním jsou přiřazovány všechny možné systémové interpretace, které se stanou předmětem manuální desambiguace. Takto připravená zdrojová data poslouží pro natrénování taggeru umožňujícího automatickou desambiguaci. Tagger s takto ošetřenými daty bude možné využít pro další připravované diachronní korpusy.

7. DESAMBIGUACE

Pro manuální desambiguaci morfologicky oanotovaných zdrojových dat se využívá nástroj FEAT,[28] jehož podoba byla upravena tak, aby umožňovala zrealizovat naše pojetí desambiguace jazyka 19. století.[29] Tento nástroj nabízí anotátorovi všechny možné interpretace pro morfologickou anotaci, z kterých anotátor vybere jednu adekvátní, popřípadě může vytvořit vlastní, pokud by byly všechny z nabízených interpretací

25 Seznamy paradigmat jsou nagenerovány ve formátu txt a kódování UTF-8. Při jejich kontrole bylo možné využít editor Excel.

26 Týká se to lexémů obsahujících ve svém paradigmatu nestandardní tvar, který není typický pro zvolený vzor, v jehož rámci byly nagenerovány.

27 V této fázi jsou již zdrojové texty ve formátu, který lze využít pro zachycení struktury korpusu a textů v něm, v tzv. vertikále. V rámci vertikály jsou zdrojová data pomocí tokenizace rozčleněna na jednotlivé tokeny, k nimž je posléze přiřazena morfologická charakteristika.

28 Viz *Feat / Home* [online]. [cit. 29. 3. 2022]. Dostupné z: https://bitbucket.org/czesl/feat/wiki/Home>.

29 Např. možnost zachycení vícečlenného lemmatu u tvaru *zemí: zem/země N-FP2-----A----- N-FS7-----A-----*.

nevyhovující.[30] Anotace je prováděna paralelně, tzn. že dva anotátoři anotují totožnou část zdrojových dat vždy proti sobě (paralelně) a posléze je verifikována mezianotátorská shoda.

8 TAGSET

U morfologické anotace korpusu DIA1900 vycházíme z pražského (Hajičova) šestnáctimístného tagu[31] užívaného pro morfologickou anotaci synchronních korpusů češtiny. Původní tagset byl pro naše účely v několika pozicích upraven, aby bylo možné 1) označit jazykové kategorie, u nichž jsme přesvědčeni, že je přínosné je zohlednit (nezávisle na jazyku 19. století), přičemž v synchronních korpusech specifikovány nejsou; 2) zachytit odlišnosti jazyka 19. století oproti současné češtině. V důsledku toho byly některé pozice z šestnáctimístného tagu modifikovány, některé pak vyjadřují zcela odlišnou jazykovou kategorii. Modifikací některých pozic je míněno jejich zjednodušení a zpřehlednění, případně doplnění.[32] Další pozice z původního Hajičova tagsetu jsou nahrazeny jinými jazykovými kategoriemi, které lépe umožní formálně specifikovat jazyk korpusu DIA1900.[33]

Zásadní změnu v rámci označení jazykových kategorií reprezentuje specifikace proprií na 6. pozici tagu (srov. dále). Jedná se o kategorii, která je jako součást mor-

30 V rámci koncepce morfologické anotace se pracuje i s tzv. vícečlenným lemmatem. Tento druh lemmatu je přičleněn ke všem tvarům, u nichž nelze jednoznačně určit jen jedno lemma z vícera možných. K takovým tvarům se řadí např. tvar *zemí*, který bude mít ve výsledném korpusu DIA1900 vícečlenné lemma *zem/země*. Vícečlenné lemma není omezeno počtem dvou lemmat, např. ke slovesnému tvaru *nalezl* bude přiřazeno vícečlenné lemma v podobě *nalézt/nalezt/nalízt/naleznout*.

31 ÚTKL: *Poziční morfologické tagy* [online]. [cit. 29. 3. 2022]. Dostupné z: <http://utkl.ff.cuni.cz/~skoumal/morfo/?lang=cs>; Hajič, J. (2004).

32 Úpravy tohoto druhu se týkají zejména druhé pozice tagsetu (slovní poddruh), která byla radikálně zjednodušena a zpřehledněna; drobnější změny byly provedeny dále na pozici první, třetí, čtvrté, páté, osmé a šestnácté (jedná se především o odstranění interpretace libovolnosti (neurčitosti) u čísla, pádu atd.).

33 Změny vyjadřovaných jazykových kategorií v původním Hajičově tagsetu byly provedeny na těchto pozicích (Hajičův tagset > **tagset pro DIA1900**): na 6. pozici: přivlastňovací rod > **proprium**, na 7. pozici: přivlastňovací číslo > **duál**, na 9. pozici: čas > **NEVYUŽITO**, na 14. pozici: NEVYUŽITO > **agregát** (SYN2020 kategorii agregátu využívá, formálně však v rámci tohoto korpusu není součástí tagu), na 15. pozici: varianta, stylový příznak > **tvary obsahující částici.** Agregáty jsou míněna slova, kterým se v procesu anotace přiřazují současně dvě (výjimečně tři) řady pozičních atributů. Důvodem je fakt, že se sice tato slova zapisují jako slovo jedno, avšak z pohledu určování gramatických kategorií se chovají jako slova dvě (tři). Z hlediska tokenizace agregáty představují jednu textovou pozici (jeden token), zatímco morfologická analýza s nimi zachází tak, jako by se jednalo o více pozic. Agregátem je např. tvar *běžels* s lemmatem *běžet_být* a morfologickou značkou *VpIS------AA-1-I_VB-S---2--AA-1-I*. Pro lepší přehlednost jsou tyto pozice v níže uvedeném tagsetu podbarveny šedě.

fologické anotace poprvé zachycena až v diachronním korpusu DIA1900; synchronní korpusy s ní nepracují. Dalším z frekventovaných jazykových jevů, který lze v porovnání se současnou češtinou charakterizovat jako odlišnost, je využívání příklonných částic.[34] Pro zachycení této jazykové reality jazyka 19. století[35] byla využita 15. pozice v tagu. S tím mimo jiné souvisí velikost MS, který se tak stal objemnějším, jelikož kompletní adjektivní paradigmata a část zájmenných a slovesných paradigmat a adverbií bylo nutné vedle paradigmat bez příklonných částic nagenerovat také s těmito částicemi.

Pro přehlednost a shrnutí je přiložen tagset pro korpus DIA1900, který má tuto podobu:

POZICE 1 (slovní druh)

C	numerál (číslovka, nebo číselný výraz s číslicemi)
A	adjektivum (přídavné jméno)
D	adverbium (příslovce)
I	interjekce (citoslovce)
J	konjunkce (spojka)
N	substantivum (podstatné jméno)
P	pronomen (zájmeno)
R	prepozice (předložka)
T	partikule (částice)
V	verbum (sloveso)
X	neznámý, neurčitelný slovní druh
Z	interpunkce
F	cizí slovo
Y	slovní druh jednoznačně neurčitelný v daném kontextu

POZICE 2 (slovní poddruh)

	A:
A	adjektivum obecné (kvalitativní nebo vztahové), složené tvary nepřivlastňovací
C	adjektivum obecné (kvalitativní nebo vztahové), jmenné tvary (mrtev)
U	adjektivum přivlastňovací (otcův, matčin, Novákovic)
	C:
l	číslovka základní vč. neurčitých
r	číslovka řadová vč. neurčitých
v	číslovka násobná vč. neurčitých
y	číslovka dílová vč. neurčitých
d	číslovka druhová, adjektivní skloňování „dvojí“ vč. neurčitých
j	číslovka druhová >= 4, substantivní postavení „čtvero“ vč. neurčitých

34 Jedná se o příklonné částice *-ť, tě* a *-ž- že*.

35 V 19. století mají příklonné částice už jen funkci zesilovacího prostředku a zdůrazňují slovo (tvar slova), jehož jsou součástí. Během historického vývoje lze však předpokládat, že příklonné částice byly polyfunkčním jazykovým prostředkem, který neměl jen zesilovací funkci.

k číslovka druhová >= 4, adjektivní postavení, krátký tvar „čtvery“, vč. neurčitých
? číslovka „kolik“
= číslo psané arabskými číslicemi (slovní druh: číslovka — ‚C‘)

D:

G spřežková (nasucho — psáno dohromady)
- ostatní

I:

- neurčuje se

J:

, spojka podřadicí
^ spojka souřadicí

N:

- ostatní (apelativum)

P:

W zájmeno záporné (nic, nikdo, nijaký, žádný, nižádný, pražádný, nijeden, ničí, nikterý, nesvůj…)
Z zájmeno (nezáporné) neurčité (něčí, číkoli, leččí, ledačí, zájmena odvozená od kdo, co, který, jaký: leckdo, leda(s)kdo, všelikdo, nějaký, kdekdo, bůhvíkdo, čertvíco, sotvakdo…)
P zájmeno (nezáporné, nikoli neurčité) osobní (já, ty on, ona, ono, my, vy, oni, ony, ona)
S zájmeno (nezáporné, nikoli neurčité) přivlastňovací „základní/nespecifické“ (můj, tvůj, svůj, jeho, její, náš, váš, jejich)
1 zájmeno (nezáporné, nikoli neurčité) přivlastňovací vztažné (jehož, jejíž, čí)
8 zájmeno (nezáporné, nikoli neurčité) přivlastňovací tázací (čí, čípak)
L zájmeno (nezáporné, nikoli neurčité, nikoli přivlastňovací) vymezovací (taký, takový, onaký, týž, tentýž, sám, každý, všechen, všecek…)
D zájmeno (nezáporné, nikoli neurčité, nikoli přivlastňovací) ukazovací (ten, tento, tenhle, tamten, tamhleten, onen…)
K zájmeno (nezáporné, nikoli neurčité, nikoli přivlastňovací) tázací (kdo, co, který, jaký, čí, kdopak, čípak…)
4 zájmeno (nezáporné, nikoli neurčité, nikoli přivlastňovací) vztažné (kdo, co, který, jaký, jenž)
6 zájmeno (nezáporné, nikoli neurčité, nikoli přivlastňovací) zvratné (se)

R:

R předložka, základní tvar
V předložka, vokalizovaný tvar
F součást předložky, která nikdy nestojí samostatně („vzhledem“…)

T:

7 reflexivní zájmenná částice (se, si, sobě)
- ostatní

V:

B slovesný tvar indikativu aktiva vyjadřující přítomnost nebo budoucnost
p slovesný tvar — příčestí minulé (rezignuje se na určování antepréterita)
s trpné příčestí
c tvary volného kondicionálového morfému bych, bysem, bys, bysi, by, bychom, bysme, byste odvozené od historických aoristových tvarů slovesa být
e přechodník přítomný
m přechodník minulý

f infinitiv
i imperativ
X:
- neurčuje se
Z:
- neurčuje se
F:
- neurčuje se
Y:
o součást (budoucí) spřežky nebo složeniny ([na] sucho; zčista jasna)

POZICE 3 (jmenný rod)
F rod ženský
M rod mužský životný
I rod mužský neživotný
N rod střední
- tvar nevyjadřuje jmenný rod (např. jsem, bych apod.)

POZICE 4 (číslo)
P plurál (množné číslo)
S singulár (jednotné číslo)

POZICE 5 (pád)
1 nominativ (1. pád)
2 genitiv (2. pád)
3 dativ (3. pád)
4 akuzativ (4. pád)
5 vokativ (5. pád)
6 lokativ (6. pád)
7 instrumentál (7. pád)

POZICE 6 (vlastní jméno)
j proprium nebo součást víceslovného vlastního jména — jména osob, zeměpisná jména, instituce
- apelativum

POZICE 7 (duál)
1 formálně duálový tvar (s koncovkou -ma, příp. tvary párových substantiv kolenou, ramenou, očí)
- neurčuje se

POZICE 8 (osoba)
1 1. osoba
2 2. osoba
3 3. osoba

POZICE 9 (pomocné sloveso)
- neurčuje se

POZICE 10 (stupeň)
1 1. stupeň
2 2. stupeň
3 3. stupeň
- neurčuje se

POZICE 11 (negace)
A afirmativ (bez negativní předpony „ne“)
N negace (tvar s negativní předponou „ne“)
- neurčuje se

POZICE 12 (aktivum/pasivum)
A aktivum nebo ‚nikoli pasivum'
P pasivum

POZICE 13 (NEVYUŽITO)
- neurčuje se

POZICE 14 (agregát)
1 součást grafické jednotky složené z tvarů několika slov
- neurčuje se

POZICE 15 (tvary obsahující částici)
T tvar obsahující částici -ť, -tě, -ž
B tvar obsahující kondicionálový morfém -by
- neurčuje se

POZICE 16 (vid)
P dokonavý vid
I nedokonavý vid

9. PROPRIA

Při sestavování MS a adaptaci tagsetu pro účely korpusu DIA1900 bylo možné označit nejen specifické morfologické, ale i sémantické charakteristiky, což umožnilo v rámci korpusů ÚČNK zcela nově identifikovat propria, konkrétně na 6. pozici tagu. Propria jsou co do své funkce i morfologické a sémantické specifičnosti svébytnou skupinou (Knappová, 1980; Hladká — Nekula, 2017), jejíž jednoznačné vydělení otevře možnosti dalšího lingvistického výzkumu. Z hlediska korpusového zpracování proprií a jejich začlenění do MS byla výzvou především morfologická specifika proprií, jejich homonymie s apelativy a celkově také detekce propriálních tvarů v textu.

Z pohledu zařazení proprií do MS jde např. o odchylky od apelativních vzorů dané absencí singulárové, případně plurálové části paradigmatu, dubletní pádové koncovky i koncovky pro odpovídající vzor nestandardní či obecně nepravidelné paradigma. V případě některých proprií cizího původu bylo zapotřebí vytvořit vzory ryze propriální a u obtížně zařaditelných proprií s malou mírou začlenění do českého pravopisného a morfologického systému se omezit pouze na označení konkrétního slovního tvaru v textu, bez generování celého paradigmatu.

Pro názornost jmenujme několik příkladů: příjmení cizího původu s němým *e* byly rozřazovány např. ke vzorům *Outsider*, *Zwinger*, *Voltaire*. K typickým dubletám doplňovaným ke vzorům u toponym patřily u pomnožných koncovky *-ům/-ím* v dativu a *-ech/-ích* v lokálu, dále např. u toponym zakončených *-burk/-purk* bylo zapotřebí nagenerovat dublety *-a/-u* v genitivu a *-ku/-ce* v lokálu apod.

Vedle formálních zvláštností propriálních tvarů bylo nutné ošetřit i homonymii s tvary apelativními, u nichž by v případě nedetekování propriální platnosti došlo k chybné morfologické interpretaci tvaru také na pozici slovního druhu či rodu.

Jde např. o typ *pan Sluníčko* (substantivum, maskulinum životné), *pan Nosil* (substantivum, maskulinum životné), *pan Lomnický* (substantivum, maskulinum životné) nebo *paní Sobolová* (substantivum, femininum). Pro příjmení s adjektivní formou tak byly vytvořeny propriální vzory *Palacký*, *Nebeský*, které vedle již existujících apelativních vzorů *Hajný* a *Kojná* umožnily substantivní interpretaci těchto propriálních tvarů.

Samotné budování modulu proprií v rámci MS bylo rozčleněno do několika fází. Propria byla do MS slovníku zařazována v rámci první fáze přípravy spolu s apelativními lemmaty ze SSJČ a PSJČ, kdy propriální platnost lemmatu signalizovalo počáteční velké písmeno. Paralelně probíhala detekce proprií v materiálu korpusu DIA1900. Propria byla dále selektována v rámci procesu ruční předanotace zaměřené na automaticky označené nestandardní tvary. Nakonec byly detekovány propriální tvary homonymní se standardními tvary apelativ. Pro tento proces bylo zásadní, aby byl MS již v závěrečné fázi příprav a obsahoval co největší množství apelativních tvarů. Automaticky byly z textového materiálu korpusu vybrány takové tvary, které měly velké počáteční písmeno, ale jinak byly formálně totožné s apelativním tvarem obsaženým v MS. Výsledný výběr tvarů byl dále tříděn a propriální tvary byly rozřazovány k adekvátním vzorům. Na závěr (po podchycení víceslovných propriálních jednotek) byl využit nástroj NameTag, detekující automaticky různé druhy pojmenování.[36]

Jako zatím neproveditelná se ukázala původní vize provázanosti variantních toponym označujících tutéž lokalitu. Vzhledem k velké škále variability, pravopisných zvláštností, neobvyklého užití, přítomnosti variant různých jazykových proveniencí přizpůsobených často těžko předvídatelným způsobem i vzhledem k obtížné identifikaci označovaných míst nebylo (zatím) možné tuto představu realizovat.

Lze však předpokládat, že se modul proprií s dalším rozšiřováním MS bude zdokonalovat a určení proprií v každém dalším korpusu s využitím tohoto slovníku bude snazší a přesnější.

36 *NameTag* [online]. [cit. 29. 3. 2022]. Dostupné z: <http://lindat.mff.cuni.cz/services/nametag/run.php>.

10. ZÁVĚR

Ačkoli jsme při tvorbě korpusu DIA1900 dbali o maximálně přesný formální popis jazyka 19. století, snažili jsme se zároveň stejnou měrou zohlednit potřeby potenciálního uživatele tohoto korpusu s jeho synchronním vnímáním češtiny. Tento záměr v důsledku určoval veškerý postup prací při přípravě korpusu. Tam, kde byla při práci s jazykovým materiálem tato dvě kritéria na první pohled neslučitelná a bylo nutné mezi nimi volit, byly vyvinuty postupy, které zajistily větší plastičnost popisu, a umožnily tak zachování rovnováhy mezi kritérii. Na jednotlivá rozhodnutí na úrovni zpracování textů v tomto smyslu navazují rozhodnutí o organizaci tvarů a lemmat MS a hledání nových technických možností, které by pomohly provázat variantní tvary a lemmata (např. volba vícečlenného lemmatu a našeptávač). Vzhledem k tomu, že příprava korpusu a metadat k jeho anotování byla podstatnou měrou ovlivněna omylností lidského faktoru (nicméně tento faktor považujeme spíše za klad našeho přístupu), jsme si při vší naší snaze vědomi, že výsledný korpus nemůže být zcela prost chyb. Rovněž pochyby, které jsme měli při rozhodování o sporných prvcích, zůstanou nadále integrální součástí korpusu DIA1900, ač primárně neviditelné a námi snad vyřešené tím nejadekvátnějším možným způsobem.

Přesto, nebo snad spíše díky tomu, že valná většina prací spojená s přípravou korpusu DIA1900 byla v základu manuální nebo poloautomatická, umožní výsledný MS spolu s ručně desambiguovanými daty v budoucnu spolehlivou automatickou anotaci textů celého 19. století, případně i textů starších.

Za využití vytvořeného MS a etalonu a za stejných kritérií týkajících se žánrové a časové vyváženosti vznikne v budoucnosti korpus 1. poloviny 19. století — DIA1850, zahrnující texty z 1. poloviny 19. stol. Jeho realizací by se nám podařilo z jazykově formálního hlediska kompletně postihnout češtinu celého 19. století, a to už s využitím automatické morfologické anotace a desambiguace. Než se tak stane, bude nutné vytvořit manuálně desambiguovaný etalon z textů 19. století, který se stane základem pro tagger.[37]

LITERATURA A INTERNETOVÉ ZDROJE

Atom. A hackable text editor for the 21st Century [online]. [cit. 29. 3. 2022]. Dostupné z: <https://atom.io/>.

DEBDict [online]. [cit. 29. 3. 2022]. Dostupné z: <https://deb.fi.muni.cz:8005/debdict/>.

Feat / Home [online]. [cit. 29. 3. 2022]. Dostupné z: <https://bitbucket.org/czesl/feat/wiki/Home>.

ÚČNK: *Wiki* [online]. [cit. 29. 3. 2022]. Dostupné z: *https://wiki.korpus.cz/doku.php/pojmy:lemma?s[]=hyperlemma.*

37 V současné chvíli (březen 2023) je dokončena většina přípravných fází korpusu DIA1900 popisovaných v tomto příspěvku. Aktuálně je prováděna ruční desambiguace etalonu, která představuje prefinální fázi při dokončování tohoto korpusu.

ÚTKL: *Poziční morfologické tagy* [online]. [cit. 29. 3. 2022]. Dostupné z: <http://utkl.ff.cuni.cz/~skoumal/morfo/?lang=cs>.

Hajič, J. (2004): *Disambiguation of Rich Inflection: Computational Morphology of Czech.* Praha: Karolinum.

Hladká, Z. — Nekula, M. (2017): VLASTNÍ JMÉNO. In: P. Karlík — M. Nekula — J. Pleskalová (eds.), *CzechEncy — Nový encyklopedický slovník češtiny.* [online]. [cit. 15. 3. 2022]. Dostupné z: <https://www.czechency.org/slovnik/VLASTNÍ JMÉNO>.

Hlaváčová, J. — Mikulová, M. — Štěpánková, B. — Hajič, J. (2019): Modifications of the Czech morphologigal dictionary for consistent corpus annotation. *Journal of Linguistics,* 70, 2, s. 380–389.

Internetová jazyková příručka [online]. [cit. 29. 3. 2022]. Dostupné z: <https://prirucka.ujc.cas.cz/>.

Kieraś, W. — Woliński, M. (2018): Manually Annotated Corpus of Polish Texts Published between 1830 and 1918. In: *Proceedings of the Eleventh International Conference on Language Recources and Evaluation (LREC 2018)*. Miyazaki, Japan: ELRA.

Knappová, M. (1980): Významové aspekty vlastních jmen. *Slovo a slovesnost,* 41, s. 57–60.

Kučera, K. — Najbrtová, K. — Pivoňková, K. — Řehořková, A. — Stluka, M. (2019): Korpus českého jazyka 2. poloviny 19. století. *Časopis pro moderní filologii,* 101, 1, s. 92–97.

Linde, S. — Mittmann, R. (2013): Old German reference corpus: digitizing the knowledge od the 19th century. Automated pre-annotation using digitized historical glossaries. In: P. Bennett — M. Durrell — S. Scheible — R. J. Whitt (eds.), *New Methods in Historical Corpora*. Tübingen: Narr Verlag, s. 235–246.

NameTag [online]. [cit. 29. 3. 2022]. Dostupné z: <http://lindat.mff.cuni.cz/services/nametag/run.php>.

Osolsobě, K. — Hlaváčová, J. — Petkevič, V. — Šimandl, J. — Svášek, M. (2017): Nová automatická morfologická analýza češtiny. *Naše řeč,* 4, s. 225–234.

Pettersson, E. — Borin, L. (2019): *Characteristics of diachronic and historical corpora. Features to consider in a Swedish diachronic corpus.* [online]. [cit. 29. 1. 2022]. Dostupné z: <https://sweclarin.se/sites/sweclarin.se/files/diachronic-corpora-sweclarin-v3.pdf>.

PSJČ. Příruční slovník jazyka českého [online]. [cit. 29. 3. 2022]. Dostupné z: <https://bara.ujc.cas.cz/psjc/>.

SSJČ. Slovník spisovného jazyka českého [online]. [cit. 29. 3. 2022]. Dostupné z: <https://ssjc.ujc.cas.cz/>.

Lucie Benešová | Ústav Českého národního korpusu,
Filozofická fakulta Univerzity Karlovy | Panská 890/7, 110 00 Praha 1
ORCID ID: 0000-0001-6395-3958
lucie.benesova@ff.cuni.cz

Karel Kučera | Ústav Českého národního korpusu,
Filozofická fakulta Univerzity Karlovy | Panská 890/7, 110 00 Praha 1
ORCID ID: 0000-0002-0762-5682
karel.kucera@ff.cuni.cz

Kateřina Najbrtová | Ústav Českého národního korpusu,
Filozofická fakulta Univerzity Karlovy | Panská 890/7, 110 00 Praha 1
ORCID ID:
katerina.najbrtova@ff.cuni.cz

Klára Pivoňková | Ústav Českého národního korpusu,
Filozofická fakulta Univerzity Karlovy | Panská 890/7, 110 00 Praha 1
ORCID ID: 0009-0000-0990-7424
klara.pivonkova@ff.cuni.cz

Martin Stluka | Ústav Českého národního korpusu,
Filozofická fakulta Univerzity Karlovy | Panská 890/7, 110 00 Praha 1
ORCID ID: 0000-0003-3294-3583
martin.stlukaa@ff.cuni.cz