ČASOPIS PRO MODERNÍ FILOLOGII 101, 2019, Č. 1, S. 92–98

OPEN ACCESS

Korpus českého jazyka 2. poloviny 19. století[1]

Karel Kučera (Praha) – Kateřina Najbrtová (Praha) –
Klára Pivoňková (Praha) – Anna Řehořková (Praha) –
Martin Stluka (Praha)

CORPUS OF THE CZECH LANGUAGE OF THE 2ND HALF OF THE 19TH CENTURY

The paper describes the principles and structure of the one-million-word DIA1900 Corpus built at the Institute of the Czech National Corpus (CNC) in Prague, focused on the language of Czech texts published in the years 1851 to 1900. The DIA1900, planned for publication by June 2020 and to be followed by the DIA1850 (a corpus built around the same principles, with the focus on the first half of the 19th century), observes both the balanced representation of the three major text types (belles lettres — journalistic texts — technical/scientific texts) and the system of morphological tagging implemented in the synchronic corpora included in the CNC project, thus facilitating the diachronic comparison of two stages in the development of Czech.

A brief description is given of the structure of the morphological terminology used in the lemmatisation and tagging of the corpus, and of two tools designed to help search the 19th century texts with their fluctuating orthographic consistency combined with phonological and morphological variation characteristics of the language of the period: (1) a multiple select/suggest feature (reminding the user of the existence of non-standard orthographic and phonological variants of the lemma found in the corpus before the lemma search is started) and (2) the position attribute (informing the user of the ambiguous status of a word in the text, resulting from a misprint or misspelling, damaged page etc.).

KEYWORDS

diachronic corpus, lemmatisation, morphological tagging, post-national revival Czech, 19th century Czech, phonological variability, orthographic variability, morphological variability

KLÍČOVÁ SLOVA

diachronní korpus, lemmatizace, morfologické značkování, poobrozenská čeština, čeština 19. století, hlásková variabilita, pravopisná variabilita, morfologická variabilita

DOI

https://doi.org/10.14712/23366591.2019.1.6

1. CHARAKTERISTIKA KORPUSU A JEHO KONCEPCE

Korpus DIA1900, který je v Ústavu Českého národního korpusu (ČNK) Filozofické fakulty UK v Praze připravován ke zpřístupnění v první polovině roku 2020, je prv-

1 Tento článek vznikl při realizaci projektu Český národní korpus (LM2015044) financovaného Ministerstvem školství, mládeže a tělovýchovy v rámci aktivity Projekty velkých infrastruktur pro VaVaI.

ním ze dvou historických korpusů českého jazyka kladoucích si za cíl svým složením i jednotnou koncepcí anotace umožnit co nejvšestrannější srovnání češtiny 19. století s češtinou současnou, tak jak je zachycena v korpusech řady SYN. Ve vztahu k tomuto obecnému cíli budou jak v rozpracovaném korpusu DIA1900, zaměřeném na jazyk 2. poloviny 19. století, tak v plánovaném korpusu DIA1850, orientovaném na češtinu 1. poloviny 19. století, dodržovány tři základní principy odpovídající principům uplatňovaným v korpusech synchronních, konkrétně

(1) jednotný rozsah korpusů (v případě korpusů DIA milion slovních tvarů),
(2) vyvážené zastoupení tří textových typů (publicistika, odborné texty a beletrie),
(3) budování korpusu z celých textů.

V budoucnu, po vybudování obou korpusů, bude možno uvažovat i o dalších korpusech, které by stejně jako DIA1900 a DIA1850 mapovaly jednotlivé stavy češtiny a její vývoj v půlstoletých obdobích směrem do vzdálenější minulosti. Vzhledem k nevyřešené legislativě týkající se copyrightu a uvolňování novějších publikovaných textů pro využití k vědeckým účelům by v současné době bylo předčasné zvažovat rozšíření záběru korpusů DIA i v opačném směru, tj. na první polovinu 20. století.

Korpus DIA1900 je složen z textů různého rozsahu (2 200–29 300 slovních tvarů) publikovaných v letech 1851–1900, a to ve výše zmíněném vyváženém zastoupení publicistiky, odborných textů a beletrie, které je dodržováno nejen v rámci celého korpusu, ale i na úrovni jednotlivých desetiletí. Data pro korpus byla vybrána z textů, které byly digitalizovány metodou OCR a jsou volně přístupné v digitální knihovně Kramerius 3 Národní knihovny ČR (viz http://kramerius.nkp.cz/kramerius/Welcome.do). Vzhledem k různé kvalitě OCR, závislé na vlastnostech předlohy i použitého softwaru, byly všechny texty porovnány se svými digitálními předlohami, manuálně opraveny a následně upraveny tak, aby svým členěním a metajazykovými daty odpovídaly formátu závaznému pro texty zařazované do korpusů ČNK.

Konkrétní jazykové zásady uplatňované při přípravě textů k využití v korpusu DIA1900 vyplývaly primárně z charakteristických rysů jazyka 2. poloviny 19. století a ze specifik jeho úzu. Čeština tohoto období se v rámci periodizace vývoje českého jazyka označuje jako **poobrozenská čeština 19. století** (srov. Kosek, 2017) a chápe se jako jedno ze čtyř vývojových období nové češtiny (obrozenská čeština — poobrozenská čeština 19. století — čeština 1. poloviny 20. století — čeština 2. poloviny 20. století), jejíž počátek se klade do poslední čtvrtiny 18. století. Toto zařazení by nemělo vyvolat dojem, že při tvorbě korpusů z celého novočeského období, tedy i z 19. století, by bylo možno beze změn nebo jen s drobnými modifikacemi využít propracovaných postupů obvyklých při výstavbě korpusů synchronních. Jak shrnujeme dále, korpus DIA1900 měl v tomto směru četná specifika jak v oblasti editace textů, respektive jejich celkové přípravy pro zařazení do korpusu, tak v oblasti jejich značkování, a specifické problémy bylo třeba řešit i při tvorbě morfologického slovníku určeného k lemmatizaci a morfologickému značkování jednotlivých slovních tvarů.

Základní diferenční jazykové rysy poobrozenských textů souvisejí se společenským a kulturním vývojem v českých zemích v 2. polovině 19. století, který ve srovnání s předchozími obdobími jednak vedl ke stále širšímu růstu uplatnění českého

jazyka v literatuře, kultuře, školství, publicistice, naučné i vědecké literatuře a na nižších rovinách administrativy, jednak směřoval — mimo jiné v souvislosti s rostoucím počtem aktivních uživatelů češtiny — ke zživotnění jazyka projevujícímu se odklonem od archaizující kodifikace kulturního jazyka předchozího obrozenského období, v němž se jako vzor primárně uplatňovala čeština humanistická.[2] Tyto zživotňující tendence ve skutečnosti měly hlubší historické kořeny. Jak konstatoval Alexandr Stich (Stich, 1991, s. 59 a 60) v souvislosti s barokem, „některé vývojové tendence měly svůj počátek právě v něm a přežily pak do obrozenského období, kde se rozvinuly plněji. […] Kromě toho barokní spisovný úzus těmto tendencím napomáhal tím, že se neuzavíral před systémovými inovačními procesy, které tehdy probíhaly v běžně mluveném jazyce. […] Všechny barokní inovace jsou motivovány snahou přiblížit spisovnou normu běžně mluvenému jazyku, a to té jeho podobě, která se neopírá o žádný přísně krajově vymezený nářeční základ (ale spíše o to, co se později pro většinu stalo obecnou češtinou).“

2. VARIABILITA ČEŠTINY 19. STOLETÍ A JEJÍ REFLEXE V KORPUSU

Paralelní existenci zmíněných tří vývojových procesů (pozvolný ústup „vyšší“ humanistické normy, zživotňování češtiny a její šíření do „vyšší“ kulturní, odborné i administrativní sféry) lze označit za hlavní faktor, který v 2. polovině 19. století vedl k rozsáhlé variabilitě na různých rovinách jazykové stavby i na různých úrovních výstavby textu, s centrem v rovině hláskové, pravopisné a morfologické. Podstatně méně častá byla dobová variabilita lexikální, resp. slovotvorná, jako např. *znalec/znatel, všecek/všecken/všechen, výhradní/výhradný, ryzí* (*ryzího, ryzímu…*)*/ryzý* (*ryzého, ryzému…*), *ďábelný/ďábelský, stěží/stíží/stěžkem, křestný/křestní* ap., a řídké, i když z dnešního hlediska někdy velmi nápadné byly také varianty, respektive nově se vyhraňující preference v oblasti syntaxe (srov. např. korpusové doklady nezvládnutých konstrukcí vznikajících zjevně ze snahy o preferované „úřední“ neosobní vyjadřování pomocí pasiva: *má plný nárok na takovouto podporu a ochranu státem, neboť i tomuto musí na tom záleženo býti, aby rolnictvo neklesalo; Místo však co by se mu byli měli dát napiti, zacpáno mu ještě i jediný otvor, kudy mu z venčí do vězení čerstvý vzduch docházel, aby mimojdoucí jeho nářek neslyšeli*). Z hlediska výstavby korpusu DIA1900, který má pouze morfologické, nikoli syntaktické a sémantické značkování, nevyžadovaly dva posledně jmenované druhy variant žádný zvláštní přístup: syntaktické varianty a preference nebyly v korpusu nijak označovány ani dále zpracovávány; výše zmíněné lexikální varianty (*znalec/znatel, všecek/všecken/všechen* atd.) byly zařazeny do morfologického slovníku jako samostatné lexémy, a pokud patřily k ohebným slovním druhům, byla pro ně vygenerována odpovídající paradigmata.

Jako příklady pravopisných variant v dobových textech — a současně jako příklady odlišnosti silně rozkolísaného pravopisného úzu poobrozenské češtiny od značně unifikovaného pravopisného úzu češtiny dnešní — lze uvést podoby *způsob/spůsob, dvadcet/dvatcet/dvacet, denník/deník, předce/přece, zrcádko/zrcátko, zýtra/zítra*

2 Srov. Šlosar (2017).

aj. Značná část pravopisných rozdílů vyplývala také ze souběžného užívání přejatých, respektive v dané době přejímaných a nejednotně pravopisně počešťovaných cizích výrazů, jako například *effektný/efektný, grammatika/gramatika, klassický/klasický, mythologie/mytologie, praecisní/precisní, klarinet/klarynet* aj.

Hláskovou variantnost v dobových textech by bylo možno exemplifikovat především značnou rozkolísaností kvantity samohlásek (*miska/míska*, *kniha/kníha, jmeno/jméno*, *bájka/bajka, namáhavý/namahavý, obleknouti/obléknouti* aj.), ale podobné případy většinou nelze spolehlivě rozlišit od rozkolísaného, respektive nesoustavného označování kvantity, tedy od jevu pravopisného. Početné případy variantnosti na rovině hláskosloví lze však uvést i odjinud, zejména z oblasti zjednodušování souhláskových skupin (*prázný/prázdný, zkřiknout/vzkřiknout, cnost/ctnost, třevo/střevo* ap.), užívání podob s provedenými i neprovedenými hláskovými změnami (*vorati/orati, oučel/účel, outok/útok* aj.) nebo podob, v nichž se vyskytovalo několik druhů hláskových rozdílů současně (srov. případy jako *osýpati/osejpati/vosejpati* nebo *ouředník/ouředlník/úřadník/ouřadník*).

Četné morfologické varianty týkající se jednotlivých tvarů, nikoli celých paradigmat[3] (srov. např. nominativ plurálu *Španělé/Španělové, koně/koňové*, lokál plurálu maskulin *vrcholech/vrcholích, chlévech/chlévích*, genitiv plurálu maskulin *spisův/spisů, kořínkův/kořínků*, variantní slovesné tvary *chtí/chtějí, jezdí/jezdějí, bere/béře, dýchá/dýše, dadí/dají, přednešen/přednesen, slyšán/slyšen* aj.) byly do poobrozenské češtiny téměř vesměs přejaty ze starších vývojových fází jazyka, a nejsou tedy jejím výhradním specifikem. Varianty tohoto typu byly začleněny do skriptů určených ke generování příslušných paradigmat a morfologických tagů. Výsledně tak ani z hlediska příslušnosti k odpovídajícím lemmatům, ani z hlediska gramatické interpretace nepředstavují pro uživatele korpusu žádné problémy, a nevyžadují proto nadstavbu ani rozšíření funkcionality současných korpusových nástrojů.

V případě výše komentovaných pravopisných a hláskových variant je situace výrazně odlišná: variabilita je tu podstatně častější než u variant morfologických, je z velké části nepravidelná, obtížně předpověditelná, pro běžného, do starší češtiny nezasvěceného uživatele využívajícího současné korpusové nástroje je na základě intuice jen stěží vyhledatelná, a tedy ve výsledku i těžko uchopitelná. Z tohoto důvodu — s cílem zachytit pravopisnou, hláskovou a zčásti i morfologickou variabilitu s co nejmenším zkreslením, bez apriorních kritérií a současně tak, aby představovala co nejmenší překážku pro uživatele — budou do vyhledávání v korpusech DIA1900 a DIA1850 postupně implementovány následující dva doplňky:[4]

1. našeptávač — program, který při zadání dotazu ve formě lemmatu upozorňuje uživatele především na pravopisné nebo hláskové varianty zadávaného lemmatu (např. při dotazu na lemma *osídlovat, kost, okno, ústa* nebo *loňský* upozorní na existenci

3 Jak bylo uvedeno výše, varianty, které se uplatňovaly v celých paradigmatech, byly do morfologického slovníku zařazovány jako samostatná lemmata.

4 Postupná implementace je v tomto případě nevyhnutelná vzhledem k tomu, že tyto doplňky vyžadují množství časově mimořádně náročné manuální práce, z jejíchž výsledků bude teprve vycházet výsledné softwarové řešení.

variant *osidlovat, kosť, vokno, ousta, lonský*); z praktických důvodů bude program upozorňovat i na časté formálně blízké a uživatelům často zcela neznámé varianty slovotvorné povahy, k jakým patří například rozsáhlá dobová koexistence adjektiv měkkého a tvrdého sklonění (*severní/severný, vzorní/vzorný, absolutní/absolutný, ryzí/ryzý*), nebo na oddělené psaní součástí dnešních spřežek typu *docela, neboli* (*do cela, nebo-li*). Našeptávač nabízí možnost výběru mezi vyhledáváním omezeným na jednotlivé varianty, které uživatel sám označí, a vyhledáváním všech nabízených variant.

2. poziční atribut — informace připojená ke konkrétnímu výskytu slovního tvaru v textu, která podobně jako poznámky v kritických edicích starších písemných památek upozorňuje na neobvyklost, resp. nejasnost tohoto tvaru, zejména je-li tato neobvyklost či nejasnost spojena s možnou nespolehlivostí jeho interpretace. Konkrétně může jít například o to, že psaná nebo tištěná podoba příslušného tvaru v textu se výrazně liší od běžných pravopisných zvyklostí (např. *sv.mikulášský* ‚svatomikulášský', *sv.-Petrský* ‚svatopetrský'), že v textu je zřejmá tisková chyba (např. *slnnce* místo *slunce*) nebo že část tvaru je nečitelná, a není tedy možné rozhodnout, zda v textu byla užita například podoba *knihkupectví* nebo *kněhkupectví*.

3. LEMMATIZACE A MORFOLOGICKÉ ZNAČKOVÁNÍ TEXTŮ

Jak bylo řečeno v úvodu tohoto příspěvku, koncepce diachronního korpusu DIA1900 počítá s lemmatizací a morfologickým značkováním. Morfologický slovník, který byl k tomuto cíli vytvořen, vznikl běžnou metodou, jejíž užití pro obrozenskou a poobrozenskou češtinu bylo testováno v letech 2008–2011 v evropském projektu IMPACT[5] a jejíž aplikace je plánována nebo realizována v několika historicky orientovaných lingvistických projektech (srov. např. Synková, 2017; Tichý, 2017). Zmíněná metoda spočívá ve vytvoření hesláře slovních tvarů z dobových textů,[6] následné lemmatizaci získaných slovních tvarů a generování jejich kompletních paradigmat. V případě korpusu DIA1900 vznikl výchozí heslář slovních tvarů z Jungmannova *Slovníku česko-německého*, *Příručního slovníku jazyka českého* (PSJČ) a z historicky příznakových hesel a variant hesel *Slovníku spisovného jazyka českého* (SSJČ) označených křížkem (s významem ‚zastaralý výraz') nebo zkratkami *zast.* (rovněž s významem ‚zastaralý výraz'), *dř.* (‚dříve'), *dř. ps.* (‚dříve psáno') a *nář.* ‚nářeční výraz, tvar'.[7] Tento slovníkový základ

5 Viz http://www.impact-Project.eu/uploads/media/IMPACT_DEE3.13_Czech_Lexicon_Documentation.pdf.

6 K tvorbě hesláře lze s výhodou využít i soudobých slovníků, pokud existují, ale je třeba je doplnit o jazyková data získaná z autentických dobových textů, neboť při tvorbě slovníkových děl byly z puristických či jiných pohnutek často vypouštěny i výrazy, které byly součástí běžného úzu. V češtině 19. století jde například o četná slovesa zakončená v infinitivu na *-írovat/-ýrovat*, která do češtiny pronikala z němčiny.

7 PSJČ a SSJČ byly za základ hesláře zvoleny především se zřetelem k rozsáhlé excerpci literárních a vědeckých textů 19. století, pořízené pro vydání těchto slovníků. Hesla a varianty hesel označované v SSJČ zkratkou *nář.* byly do hesláře zařazeny vzhledem

byl doplněn o nejčastější lemmata získaná poloautomatickým výběrem ze souboru dostupných elektronických textů publikovaných v 19. století. Lemmata zastoupená v takto rozšířeném heslář byla následně rozčleněna podle slovních druhů a v případě lexémů patřících k ohebným slovním druhům byla dále přiřazena buď k některému z deklinačních či konjugačních typů, anebo začleněna do souboru lexémů s nepravidelnou flexí (osobní zájmena, slovesa jako *být*, *mít* ap.), jejichž paradigmata byla vytvářena ručně.[8]

Paradigmata k lemmatům s pravidelnou deklinací nebo konjugací byla generována souborem skriptů, které každému generovanému tvaru současně přiřazují lemma a morfologickou interpretaci ve formě šestnáctimístného tagu užívaného v korpusech řady SYN. Morfologický slovník bude poprvé využit k lemmatizaci a morfologickému označkování výběru z korpusových textů o celkovém rozsahu půl milionu slovních tvarů, který po ruční desambiguaci bude sloužit jako trénovací datový soubor pro morfologický tagger MorphoDiTa.[9] Před samotnou lemmatizací a morfologickým značkováním celého korpusu byl v době vzniku tohoto příspěvku (duben 2019) morfologický slovník stále ještě doplňován o další apelativní lexémy z textů 2. poloviny 19. století, zejména o výrazy z dobových novin a časopisů, které do značné míry stály na okraji pozornosti obrozenských i pozdějších lexikografů. Současně byl průběžně rozšiřován i samostatný slovníkový modul dobových proprií.

LITERATURA A INTERNETOVÉ ZDROJE

Bláha, O. (2016): *Poznámky k morfologickému vývoji češtiny*. Olomouc: Univerzita Palackého v Olomouci, 2016.

Kosek, P. (2017): Periodizace vývoje češtiny. In: P. Karlík — M. Nekula — J. Pleskalová (eds.), *CzechEncy — Nový encyklopedický slovník češtiny*. URL: https://www.czechency.org/slovnik/PERIODIZACE VÝVOJE ČEŠTINY.

Stich, A. (1991): O počátcích moderní spisovné češtiny. *Naše řeč*, 74, s. 57–62.

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

Šlosar, D. (2017): Poobrozenská čeština 19. stol. In: P. Karlík — M. Nekula — J. Pleskalová (eds.), CzechEncy — Nový encyklopedický slovník češtiny. URL: https://www.czechency.org/slovnik/POOBROZENSKÁ ČEŠTINA 19. STOL.

Tichý, O. (2017): Nástroj na tvaroslovnou analýzu staré angličtiny. *Časopis pro moderní filologii*, 99, 1, s. 40–54.

k tomu, že až na novější výjimky jde o výrazy excerpované z klasických děl české literatury 19. století.

8 Analogický postup (včetně ručního vytváření nepravidelných paradigmat) je užíván při tvorbě podobných heslářů nejen pro češtinu, ale i pro další jazyky; srov. Tichý (2017), s. 46.

9 Pro podrobnější informace o tomto taggeru viz stránky Ústavu formální a aplikované lingvistiky Matematicko-fyzikální fakulty UK (http://ufal.mff.cuni.cz/morphodita).

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
katerina.najbrtova@ff.cuni.cz

Klára Pivoňková | Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy |
Panská 890/7, 110 00 Praha 1
klara.pivonkova@ff.cuni.cz

Anna Řehořková | Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy |
Panská 890/7, 110 00 Praha 1
ORCID ID: 0000-0002-6676-317X
anna.rehorkova@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
karel.kucera@ff.cuni.cz