HISTORICKÝ KORPUS SLOVENČINY

Radovan Garabík
Jazykovedný ústav Ľudovíta Štúra SAV
Bratislava

GARABÍK, Radovan: Corpus of Historical Slovak. Slovak Language, 2019, Vol. 84, No 3, pp. 307 – 317.

Abstract: The article presents the structure of the Corpus of Historical Slovak – a diachronic corpus of written Slovak texts predating language standardization attempts (texts from the 15th to the 18th century). The content of the corpus is based predominantly on existing published transcribed manuscripts, in this sense it is an opportunistic corpus, aiming to collect primarily existing texts; but we also collect and transcribe some documents directly, in order to improve the chronological balance of the corpus. The corpus aims for historical accuracy captured orthography-wise, but given existing standards in transcribing historical Slovak, this was not always possible with complete accuracy.

Keywords: corpus, diachronic, history, Slovak language

ÚVOD

Historický korpus slovenčiny je diachrónnym korpusom písanej slovenčiny predspisovného obdobia, t. j. obsahuje texty zhruba medzi 15. a 18. storočím. Tento článok opisuje koncepciu korpusu, jeho anotáciu, štruktúru a používanie (pri ktorom sa predpokladá základná orientácia v používaní korpusového manažéra NoSketch Engine (Rychlý 2007)). Článok sa úmyselne nevenuje metodike výberu textov ani spracovaniu konkrétnych textov, podobne sa nesnaží opísať zdroje vstupujúce do korpusu (ich podrobná historická a jazyková analýza je v konkrétnych prípadoch spracovaná v literatúre uvedenej v *Bibliografii zdrojov textov v korpuse* na konci tohto článku).

Na úvod v kontexte okolitých slovanských jazykov spomenieme niektoré významné diachrónne korpusy, ktoré čiastočne slúžili ako ponaučenie pri návrhu a ďalšom rozvoji historického korpusu slovenčiny. Korpus slovenčiny bol pritom na jednej strane obmedzený limitovanými možnosťami a zdrojmi, čo sa prejavilo na pragmatickom spracovaní textov, ktoré už boli prevažne transliterované a vydané, alebo priamo dostupné v elektronickej podobe. Na druhej strane sa autori korpusu neorientovali na samostatný vedecký výskum v oblasti historickej lingvistiky, ale skôr na spracovanie existujúcich materiálov. Napríklad veľmi kvalitne teoreticky spracova-

ný pojem *hyperlemy* v českom historickom korpuse (Kučera 2007) je bezproblémovo aplikovateľný aj na slovenčinu, avšak jeho uvedenie do praxe v českom korpuse je problematické a v slovenskom prostredí by bolo prakticky nemožné, vzhľadom na veľké požiadavky na manuálnu anotáciu. Slovinský korpus je zaujímavý existenciou ručne označkovaného podkorpusu, ktorý je možné použiť na trénovanie nástrojov na automatické spracovanie jazyka (Scherrer, Erjavec 2013). Návrh a tvorba takéhoto manuálne anotovaného korpusu by bola možná aj v slovenskom prostredí, avšak ako samostatný projekt. Rozsah poľských diachrónnych korpusov zase ťaží z väčšieho množstva textov v historickom období a zo štandardizovanejšej ortografie. V ďalšom texte sa pod transliteráciou rozumie prepis pôvodných textov so zachovaním ortografických vlastností a charakteru pôvodného zápisu nahradením pôvodných písmen písmenami cieľovej abecedy (hoci nie nevyhnutne reverzibilným; napríklad dlhé *ſ* môže byť transkribované písmenom *s*, keďže býva považované iba za typografický variant toho istého písmena); naproti tomu transkripcia je prepis podľa zásad cieľového pravopisu (môžu sa však pritom zachovať niektoré vlastnosti pôvodného pravopisu).

Diachrónny korpus češtiny

Diachrónny korpus Ústavu Českého národního korpusu *Diakorp*, aktuálne vo verzii 6 (Kučera, Řehořková, Stluka 2015), obsahuje okolo 4.1 milióna tokenov. V korpuse sú texty od 14. storočia po začiatok dvadsiateho storočia, ktoré sú transkribované podľa zásad moderného českého pravopisu (s výnimkami), nie transliterované. Korpus nie je lematizovaný ani morfologicky označkovaný; verejnosti je prístupný[1] prostredníctvom korpusového manažéra *Kontext*.

Diachrónne korpusy poľštiny

V poľskom korpusovolingvistickom prostredí je oblasť diachrónnej korpusovej lingvistiky a diachrónnych korpusov živá a dynamická. Urobiť úplný a detailný zoznam prístupných korpusov by bolo nad rámec tohto článku, preto uvádzame len niekoľko z nich:

- Elektroniczny Korpus Tekstów Polskich XVII i XVIII w. vznikol v Instytucie Języka Polskiego PAN[2] v Pracowni historii języka polskiego XVII i XVIII wieku[3] v spolupráci s Zespołem Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki PAN[4]. Korpus (Adamiec, 2015), obsahuje texty v dvoch ortografických rovinách – transliterácii aj transkripcii, s automatickou lematizáciou

[1] https://korpus.cz
[2] Inštitút poľského jazyka Poľskej akadémie vied
[3] Oddelenie histórie poľského jazyka XVII a XVIII storočia
[4] Skupina lingvistického inžinierstva v Inštitúte základov informatiky Poľskej akadémie vied

a morfosyntaktickým značkovaním. Korpus je dostupný[5] prostredníctvom vlastného korpusového manažéra MTAS (používajúceho jazyk CQL). Veľkosť korpusu je 12 miliónov tokenov.

- Korpus tekstów staropolskich Instytutu języka polskiego PAN (korpus nemá žiadnu súvislosť s predchádzajúcim napriek rovnakému pracovisku) obsahuje texty (starej) poľštiny do roku 1500 a je pomerne nezvykle prístupný[6] v podobe súborov v PDF alebo XML formáte s možnosťou ich stiahnutia, texty nie sú spracované v korpusovom manažéri. Transkribované sú do podoby moderného poľského pravopisu (s výnimkami) s malým množstvom interlineárnych ortografických značiek. Veľkosť korpusu je 620-tisíc tokenov.
- *ChronoPress* je korpus fragmentov novinových textov z rokov 1945 – 1954, prístupný[7] prostredníctvom vlastného rovnomenného korpusového manažéra. Texty boli starostlivo vybrané tak, aby vytvorili reprezentatívny korpus s rovnomerným pokrytím jazyka danej doby.

Diachrónne korpusy slovinčiny

Slovinský korpus *IMP* (Erjavec 2012) je založený na digitálnej knižnici starej slovinčiny od 16. storočia do roku 1918, s prevažnou väčšinou textov z druhej polovice 19. storočia. Veľkosť korpusu je 17.7 milióna tokenov, je založený na transliterácii, pričom obsahuje aj normalizovaný tvar slova (podľa moderného pravopisu). Korpus je lematizovaný a morfologicky v súlade so slovinskou korpusovou tradíciou obsahuje dve bijektívne množiny morfosyntaktických značiek: založené na slovinských gramatických termínoch a na anglických gramatických termínoch. Základom automatickej anotácie tohto korpusu je manuálne anotovaný reprezentatívny referenčný korpus *goo300k* (Erjavec 2015) s veľkosťou 358-tisíc tokenov. Všetky tieto zdroje sú prístupné prostredníctvom korpusového manažéra NoSketch Engine, ale, čo je pomerne neobvyklé, všetky dáta sú priamo stiahnuteľné.

Diachrónne korpusy slovenčiny

Dva slovenské korpusy, dostupné v rámci prístupu k Slovenskému národnému korpusu prostredníctvom korpusového manažéra NoSketch Engine, konceptuálne vychádzajú z hlavného korpusu Slovenského národného korpusu (*prim-**), a vzhľadom na jazyk sa nedajú považovať za diachrónne korpusy (ale za prekladové korpusy do „súčasnej“ slovenčiny). Napriek tomu korpusy nie sú lematizované ani morfologicky označkované.

[5] https://korba.edu.pl
[6] https://ijp.pan.pl/publikacje-i-materialy/zasoby/korpus-tekstow-staropolskich/
[7] http://chronopress.clarin-pl.eu/

Korpus[8] textov z r. 864 – 1843 vychádza zo Zlatého fondu denníka SME[9] a obsahuje texty preložené alebo prepísané do súčasnej slovenčiny (v čase vydania), často s archaickým nádychom. Veľkosť korpusu je 2.11 milióna tokenov.

Rovnako aj korpus[10] textov z r. 1843 – 1954 vychádza zo Zlatého fondu denníka SME a obsahuje texty preložené alebo prepísané do súčasnej slovenčiny (v čase vydania), často s archaickým nádychom. Veľkosť korpusu je 24 miliónov tokenov.

1. O HISTORICKOM KORPUSE SLOVENČINY

Historický korpus slovenčiny[11] obsahuje jazyk predspisovného obdobia, t. j. texty zhruba medzi 15. a 18. storočím. Korpus je koncipovaný ako oportunistický, založený (z pragmatických príčin) na už existujúcich, spracovaných prepisoch textov; napriek tomu sme pristúpili v niektorých ojedinelých prípadoch k priamemu prepisovaniu rukopisných zdrojov, na doplnenie korpusu a na zachytenie významných a známych historických zdrojov (napríklad bol prepísaný list zbojníkov mestu Bardejov z r. 1493).

Predzvesťou korpusu bolo už spracovanie Bernolákovho *Slowára Slowenského Češko-Laťinsko-Ňemecko-Uherského seu Lexici Slavicum Bohemico-Latino-Germanico-Ungaricum* (Garabík, Kajanová 2012) pre počítačové slovníkové spracovanie; niektoré praktiky a skúsenosti získané pri digitalizácii tohto diela boli potom využité pri spracovaní opísaného historického korpusu. Spracovanie zdrojov tvoriacich verziu 1.0 historického korpusu slovenčiny je opísané v Garabík – Kajanová, 2015.

Aktuálna sprístupnená (k času písania tohto článku) verzia korpusu *hist-4.0* má rozsah 917 586 tokenov. Novšia, zatiaľ nesprístupnená verzia 4.1 bola vytvorená 28. 11. 2016, opravuje drobné chyby v anotácii, odstraňuje duplicity v niektorých textoch a má rozsah 915 097 tokenov. Korpus si rozhodne nekladie za cieľ byť reprezentatívnym pre slovenčinu určitého časového obdobia, ani nechce slúžiť ako autoritatívna databáza písaných slovenských textov predspisovného obdobia, ale môže slúžiť ako referenčný korpus zachytávajúci hlavne ortografický vývoj slovenčiny predspisovného obdobia a ako odrazový mostík pre diachronický výskum gramatických a lexikálnych vlastností historickej slovenčiny.

Chronologicky prvé dokumenty v korpuse sú z roku 1451 (*Žilinská právna kniha*) a posledné z roku 1795. Okrem toho je v korpuse niekoľko dokumentov z rokov 14xx a z 17xx (určených len s presnosťou na storočie), takže teoreticky sa v ňom môžu nachádzať aj staršie či trochu novšie texty.

[8] https://korpus.sk/old1.html

[9] https://zlatyfond.sme.sk/

[10] https://korpus.sk/old2.html

[11] V skutočnosti to nie je celkom správny názov, keďže tu nejde o korpus slovenčiny, ktorý je historický, ale o korpus historického slovenského jazyka.

2. ŠTRUKTÚRA KORPUSU

2.1. Ukladanie dát

Adresárová štruktúra je v hrubých črtách inšpirovaná „najlepšími praktikami“ opísanými v štúdii R. Garabíka (2004). Dôraz bol kladený hlavne na prehľadnosť a jednoduchosť zapisovania nových dokumentov a editovania ich metadát, nie na možnosť efektívneho kolaboratívneho editovania veľkého množstva rôznorodých súborov, keďže vzhľadom na rozsah korpusu sa takéto kolaboratívne editovanie nepredpokladalo.

Texty sú v korpuse ukladané v tzv. *archíve*, tvorenom jedným adresárom. Každá jedna položka v archíve môže byť tvorená viacerými fyzickými súbormi a viacerými logickými položkami (nazývanými pre naše účely „dokumentmi“); korešpondencia medzi nimi nemusí byť koncepčne bijektívna, avšak pre prehľadnosť sme sa snažili, aby mapovanie medzi nimi bolo najviac 1:M (t. j. jednému fyzickému súboru v archíve zodpovedá jeden alebo viac dokumentov).

2.2. Anotácia metadát

Pri anotácii metadát sme sa značne odchýlili od postupov zaužívaných pri anotácii iných (primárne väčších a súčasnejších) korpusov. Dôvodom boli čisto pragmatické okolnosti, preferujúce neoddeliteľnosť textu dokumentu a k nemu prislúchajúcich metadát. Menšia veľkosť súborov a menší podiel automatizácie (oproti „bežným“ korpusom) tiež napomohli tento prístup.

Metadáta sme ukladali na začiatok každého dokumentu (to znamená, že v prípade viacerých dokumentov v rámci jedného súboru sú metadáta súčasne oddeľovačom dokumentov), kvôli uľahčeniu ďalšieho spracovávania sme každý riadok metadát uviedli znakom @ (U+0040 COMMERCIAL AT) (ktorý sa v historických textoch nevyskytuje). V prípade zápisu iba niektorých položiek v dokumentoch v rámci jedného súboru sa hodnoty nezapísaných položiek chápu ako identické s predchádzajúcimi dokumentmi – tento spôsob veľmi uľahčil anotáciu, keďže jednotná anotácia zdroja mohla byť zapísaná na začiatku súboru a k dokumentom sa zapisovali len zmenené hodnoty anotácie. Metadáta sú zapísané v jednoduchom formáte *kľúč: hodnota*, kde kľúč patrí do fixnej množiny a hodnota je v niektorých prípadoch presne formalizovaná. Metadáta sú od vlastného textu oddelené prázdnym riadkom. Ide o spôsob, ktorý dokážu bez väčšej prípravy používať aj anotátori bez skúseností s tvorbou textových korpusov.

V anotácii dokumentov boli použité tieto kľúče:

- *nr* – jednoznačný identifikátor dokumentu v rámci jednej položky v archíve, krátky reťazec zložený z alfanumerických ASCII znakov, obvykle mnemotechnická skratka názvu spracovávaného zdroja alebo poradové číslo dokumentu, ak sa zdroj skladá z viacerých dokumentov;

- *place* – geografické miesto, v ktorom dokument vznikol (obvykle názov mesta). Uvádza sa súčasný názov alebo historický názov, ak bol takto uvedený v pôvodnom dokumente;
- *date* – dátum vzniku dokumentu vo formáte ISO 8601; formálne sme stanovili, že dátumy zapisujeme v proleptickom gregoriánskom letopočte;
- *name* – výstižný, informatívny názov dokumentu, môže byť buď v originálnom jazyku (ak existuje v pôvodnom zdroji), t. j. v historickej slovenčine alebo často aj v latinčine, alebo môže byť tvorený popisom v modernej slovenčine;
- *orig* – pôvodný zdroj dokumentu, obvykle názov archívu, zbierky, fondu, knižnice;
- *comment* – prípadný komentár relevantný k danému dokumentu, často upresňujúci bibliografické zdroje alebo spôsob nadobudnutia či vzniku dokumentu.

Príklad anotácie dokumentu v archíve:

```
@nr: 2.
@place: Cífer
@date: 1599
@name: Zápis cíferského richtára o nactiutŕhaní Šimona Mestera.
@orig: Fond Magistrátu mesta Trnavy - Acta iuridica, 1551-1643,
       1599-01-08, kartón č. 1, bez sign.
@comment:
```

2.3. Spracovanie textov

Pri prepisoch textov sme sa snažili o čo najvernejšie zachovanie ortografických vlastností, pričom prednosť dostal sémantický charakter znakov pred ortografickým, avšak vždy v rámci konkrétnej ortografickej sústavy. Konkrétne ide o zachovanie grafém *g*, *j*, *w* reprezentujúcich fonémy /j/, /i:/, /v/.

Tabuľky boli zachytené v čistej textovej podobe so stĺpcami oddelenými znakom tabulátora (U+0009) a potom boli spracovávané ako lineárny text.

Z pragmatických dôvodov sme znak ‿ (U+203F UNDERTIE) zapisovali ako _ (U+005F LOW LINE) a zaviedli sme konvenciu pre zápis ľubovoľných Unicode znakov z BMP v tvare «U+NNNN+» a znakov z ostatných rovín v podobe «U+NNNNN+». Texty sú ukladané v NFKC normalizácii, pričom znak ſ (U+017F LATIN SMALL LETTER LONG S) sa už priamo v zdrojových textoch prepisoval znakom s (U+0073 LATIN SMALL LETTER S).

Počas anotácie dokumentov sme v textoch vyznačili cudzojazyčné úseky (aj na úrovni slov, ak slovo nebolo v poslovenčenom tvare použité ako súčasť textu), štruktúrnou značkou `<langXXX>`, kde `XXX` je trojpísmenový kód jazyka podľa normy ISO639-2. Celkovo sa v korpuse nachádza 221 takýchto slov alebo fráz, všetky výlučne v latinčine.

2.4. Spracovanie a prezentácia metadát

Pri tvorbe korpusu sa kládol veľký dôraz na intuitívnosť používania metadát a vyhľadávania podľa nich. Preto boli pri tvorbe anotačných metadát korpusu pôvodné metadáta skonvertované do inej podoby, lepšie vyhovujúcej potrebám používateľov korpusu.

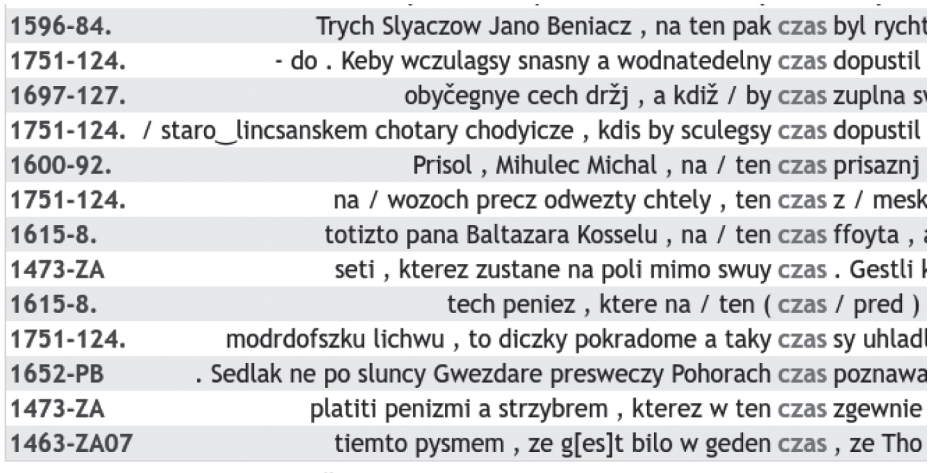
Kľúče *name*, *date* a *place* zostávajú nezmenené, kľúč orig je premenovaný na *source*. Do hodnoty kľúča *comment* boli v niektorých prípadoch doplnené informácie užitočné pre používateľa, napríklad čísla strán v *Žilinskej právnej knihe* podľa publikácie R. Kuchara (2009). Anotačný kľúč *id* bol skonštruovaný nasledujúcim spôsobom:

- prvé dva alebo štyri znaky sú skopírované z položky *date*, zodpovedajú teda roku vzniku textu, prípadne storočia-1 (ak nie je rok presne známy)[12], tento reťazec je zarovnaný sprava znakmi x na celkovú dĺžku štyroch znakov (takže sa buď rovná roku, alebo má tvar „1Nxx“, kde 1N je storočie-1);
- nasleduje znak - (U+002D HYPHEN-MINUS);
- nasleduje mnemotechnická skratka zdroja alebo súboru zdrojov (napr. ZA – *Žilinská právna kniha*, WU – *Weselé Učinki, a Rečeňj* (J. I. Bajza), UR – *Urbáre feudálnych panstiev na Slovensku*, LGS – *Vocabula latina-germanica-sclavonica* atď.); z historických dôvodov je táto časť prázdna pri dokumentoch z *Prameňov k dejinám slovenčiny*;
- nasleduje identifikátor dokumentu v rámci jedného zdroja, typicky poradové číslo (v prípade dokumentov z *Prameňov k dejinám slovenčiny* je posledným znakom tejto skratky bodka).

Príklady identifikátorov sú: 1665-UR, 1688-UR, 17xx-128., 1661-26., 16xx--LGS, 15xx-IK, 1795-WU, 1786-J2, 1786-J2M, 1779-SM, 1755-WS2, 1755-WS.

V štandardnom nastavení korpusového manažéra sa tento identifikátor zobrazuje v konkordancii ako referencia, je preto dôležité, aby prinášal čo najrelevantnejšie informácie a nebol zbytočne dlhý. Zobrazenie časového údaja je na začiatku identifikátora je v takomto type korpusu veľmi dôležité, preto figuruje na takomto prominentnom mieste a umožňuje jediným pohľadom ohodnotiť výsledky vyhľadávania v konkordancii s ohľadom na ich diachrónnosť.

[12] V ďalšom texte a v prílohách budeme na zjednodušenie odteraz používať termín „storočie“ na označenie úseku rokov začínajúcich rokom deliteľným 100 a končiacim rokom, ktorý má po delení 100 zvyšok 99. Pod termínom „storočie-1“ budeme rozumieť číslo zodpovedajúce celočíselnému deleniu roka číslom 100. Podobne pod termínom „dekáda“ budeme rozumieť označenie úseku rokov začínajúcich sa rokom deliteľným 10 a končiacich sa rokom, ktorý má po delení 10 zvyšok 9.



Obrázok 1: Ukážka vyhľadávania v korpuse. Vidno významnú úlohu referencie v rýchlom orientovaní sa v získanej konkordancii; súčasne sú tu ilustrované niektoré črty zápisu textu (rekonštrukcia slov v hranatých zátvorkách [], oddeľovač slov /, spájanie znakom ‿).

3. VYHĽADÁVANIE A ZLOŽENIE KORPUSU

3.1. Štruktúry a ich atribúty

Korpus je prístupný prostredníctvom korpusového manažéra NoSketch Engine, čo má v niektorých prípadoch vplyv na jeho dizajn a realizáciu. Korpus má hierarchickú štruktúru, na najvyššej úrovni je štruktúra `<doc>`, ktorá zároveň obsahuje ako atribúty metadáta dokumentu. Táto štruktúra zodpovedá (podľa očakávania) dokumentu v archíve. Hierarchicky priamou podradenou štruktúrou je `<p>`, zodpovedá odsekom textu (tak, ako sú zapísané alebo rekonštruované v zdrojových textoch – v originálnych textoch často odseky nebývali zaznamenané). Táto štruktúra nemá atribúty. Ďalšou štruktúrou je `<s>`, zodpovedajúca vetám. Vety boli segmentované heuristickým algoritmom založenom na interpunkcii a kapitalizácii súčasnej slovenčiny, čo zodpovedá úzu používanému v súčasných vydaniach prepisov manuskriptov a historických diel. Poslednou štruktúrou je `<g/>`. Táto štruktúra nemá otváracie a zatváracie značky, ale vyskytuje sa len na svojej pozícii v texte a indikuje, že medzi predchádzajúcim a nasledujúcim tokenom nebol biely znak.

3.2. Úroveň tokenov

Texty v Historickom korpuse nie sú lematizované ani morfologicky anotované, vyhľadávať sa dá v rovine tvarov slov a CQL s využitím regulárnych výrazov, ktoré do istej miery dokážu zastúpiť lematizáciu.

Tokenizácia textov je založená na maximalistickom princípe, tokenizované je všetko, čo sa dá rozumne tokenizovať. Toto má primárny vplyv na tokenizáciu viac-

slovných elementov spojených spojovníkom/pomlčkou alebo v historických textoch znakom = (U+003D EQUALS SIGN) (ktorý tu má sémantickú platnosť spojovníka). Tieto elementy sú vždy tokenizované na jednotlivé časti, pričom znaky spojovník/pomlčka/rovná sa sú považované za samostatný token. Podobne je aj znak _ (U+005F LOW LINE) tokenizovaný samostatne a je vo výslednom korpuse nahradený znakom ‿ (U+203F UNDERTIE) a zápisy Unicode znakov «U+NNNN+» alebo «U+NNNNN+» sú nahradené priamo týmito znakmi.

Rekonštruované vynechané časti slova (alebo sigla) sa uzuálne zapisujú v hranatých zátvorkách; tieto časti ponechávame v rekonštruovanom tvare aj s hranatými zátvorkami ako súčasť tokenu.

V korpusovom manažéri reinterpretujeme atribút *lemma*[13] a používame ho na zaznamenanie normalizovaného tvaru slova. Normalizácia pozostáva z prevodu na malé písmená a odstránenia hranatých zátvoriek, takže výsledkom je „intuitívny" tvar slova a pri vyhľadávaní nie je potrebné komplikovať regulárne výrazy možnými výskytmi hranatých zátvoriek.

ZÁVER

Historický korpus slovenčiny môže byť cenným prínosom pre výskumy jazyka v diachrónnom ponímaní. Hoci svojím rozsahom nenahrádza a nemôže nahradiť cieľavedomý a sústredený výskum primárnych zdrojov, jeho prístupnosť a spôsob spracovania môže poslúžiť pri zbežnom overovaní si predstáv o historickom vývoji slovenčiny, hlavne z hľadiska ortografických zmien. Ako perspektívu pre ďalší rozvoj korpusu by sme videli orientáciu na vylepšenie možností vyhľadávania, napríklad čiastočnou normalizáciou ortografie (čo síce nenahrádza lematizáciu či dokonca hyperlematizáciu, ale uľahčuje získavanie konkordancií), a pokračovanie v získavaní historických textov, ktoré už boli publikované v transliterovanej podobe.

Poďakovanie patrí Michaele Majerčíkovej za transliteráciu historických textov, za pridávanie anotácií dokumentov a za opravy a úpravy zdigitalizovaných dokumentov a Márii Šimkovej za myšlienku vytvorenia historického korpusu, administratívne práce s ním spojené a za konzultácie k prepisom a anotáciám textov v korpuse.

Literatúra

ADAMIEC, Dorota: Kryteria doboru tekstów do „Elektronicznego korpusu tekstów polskich z XVII i XVIII w. (do 1772 r.)". In: Prace Filologiczne, 2015, roč. 67, s. 11 – 20.

ERJAVEC, Tomaž: The goo300k corpus of historical Slovene. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Eds. N. Calzolari – K. Choukri

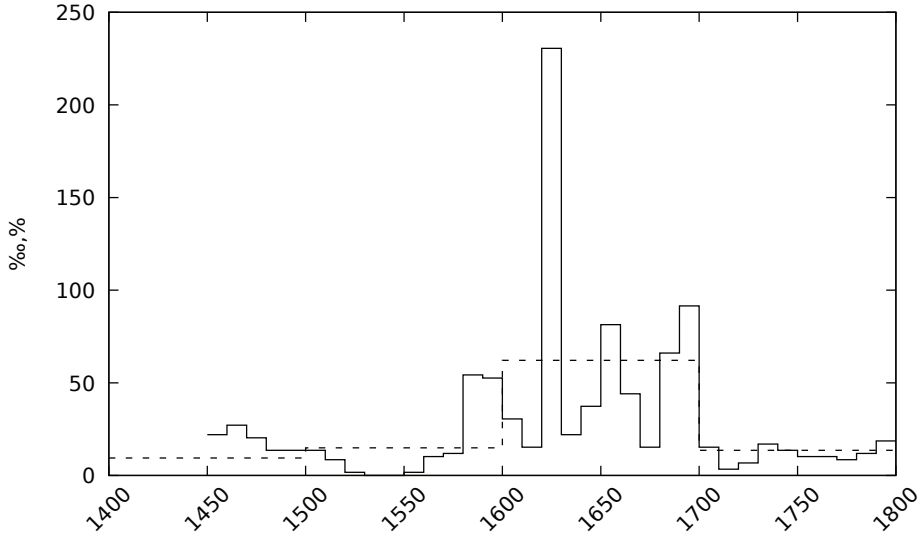
[13] Atribút by sme mohli nazvať aj iným, vhodnejším názvom. Avšak názov lemma má v NoSketch Engine špeciálne postavenie (je napríklad použitá pri konštrukcii vyhľadávania v Jednoduchom vyhľadávaní), ktoré ho robí vhodným pre používanie pri vyhľadávaní v korpuse.

– T. Declerck – M. U. Doğan – B. Maegaard – J. Mariani – A. Moreno – J. Odijk – S. Piperidis. Istanbul: European Language Resources Association (ELRA) 2012, s. 2257 – 2260.
ERJAVEC, Tomaž: The IMP historical Slovene language resources. In: Lang Resources & Evaluation, 2015, roč. 49, č. 3, s. 753 – 775.
GARABÍK, Radovan: Štruktúra dát v Slovenskom národnom korpuse a ich vonkajšia anotácia. In: Slovenčina na začiatku 21. storočia. Ed. M. Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, s. 164 – 173.
GARABÍK, Radovan — KAJANOVÁ, Michaela: Problémy a výsledky počítačového spracovania diela «Slowár Slowenskí Češko-Laťinsko-Ňemecko-Uherskí seu Lexicon Slavicum Bohemico-Latino--Germanico-Ungaricum». In: Slovo v slovníku. Aspekty lexikálnej sémantiky – gramatika – štylistika (pragmatika). Eds. K. Buzássyová – B. Chocholová – N. Janočková. Bratislava: Veda 2012, s. 294 – 300.
GARABÍK, Radovan — KAJANOVÁ, Michaela: Digitalizácia a anotácia Prameňov k dejinám slovenčiny. In: Jazykovedné štúdie XXXII. Prirodzený vývin jazyka a jazykové kontakty. Eds. K.Balleková – G. Múcsková – Ľ. Králik. Bratislava: Veda 2015, s. 577 – 583.
ISO 639-2:1998: Codes for the representation of names of languages–Part 2: Alpha-3 code.
ISO 8601:2004: Data elements and interchange formats – Information interchange – Representation of dates and times.
ISO/IEC 10646:2017: Information technology – Universal Coded Character Set (UCS).
KUČERA, Karel:. Hyperlemma: A Concept Emerging from Lemmatizing Diachronic Corpora. In: Computer Treatment of Slavic and East European Languages. Eds. J. Levická – R. Garabík. Bratislava: Tribun 2007, s. 121 – 125.
KUČERA, Karel – Řehořková, Anna – STLUKA, Martin: DIAKORP: Diachronní korpus, verze 6 z 18. 12. 2015. Ústav Českého národního korpusu FF UK, Praha 2015. Dostupný na: http://www.korpus.cz [cit. 18. 11. 2019].
RYCHLÝ, Pavel. Manatee/Bonito – A Modular Corpus Manager. In: 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University 2007, s. 65 – 70.
SCHERRER, Yves – ERJAVEC, Tomaž: Modernizing historical Slovene words with character-based SMT. In: Proceedings of the workshop : ACL 2013, The 4th Biennial International Workshop on Balto-Slavic Natural Language Processing. Stroudsburg: Association for Computational Linguistics 2013, s. 58 – 62.

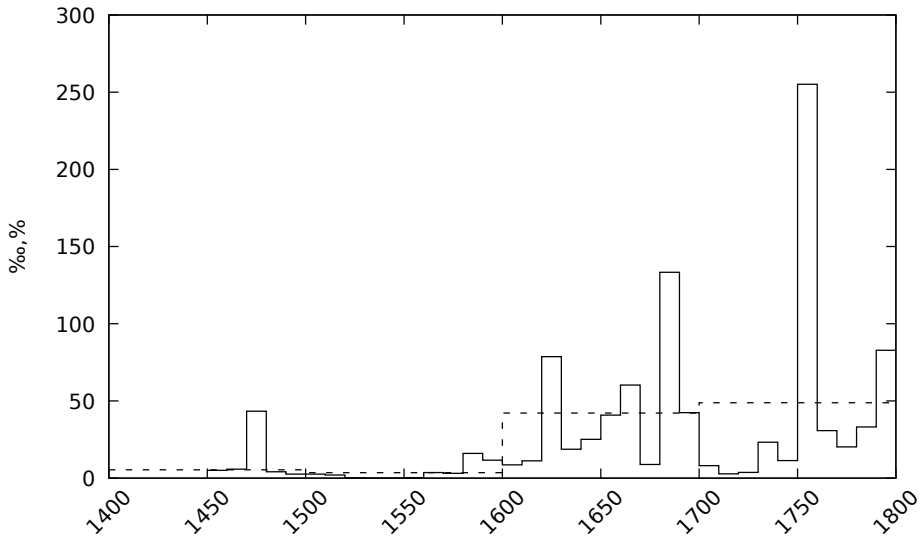
Bibliografia zdrojov textov v korpuse

KOTULIČ, Izidor: Spišský pozostalostný súpis zo 16. storočia. In: Jazykovedný časopis, 1959, roč. 10, č. 1, s. 36 – 67.
KUCHAR, Rudolf: Žilinská právna kniha. Preklad Magdeburského práva – Zápisy právnych úkonov žilinských mešťanov. Bratislava: Veda 2009. 208 s.
MATZ, S.: Scepussio-Verallaei presbyteri almae dioecesis Scepusiensis carmina. Veršovaný slovníček na konci zbierky. 1779 – 1785. In: Jazykovedný časopis, 1958, roč. 9, s. 141 – 142.
TUNYOGI, Mikuláš: Vocabula latina-germanica-sclavonica. 17. stor. In: Jazykovedný časopis, 1958, roč. 9, s. 139 – 140.
BAJZA, Jozef Ignác: Weselé učinki a rečeňj. Hač, Trnawa 1795. 364 s.
BENICKÝ, Peter: Wersse slowenské. Rkp. z r. 1652. 263 s. (UK Bratislava, O 23)
GAVLOVIČ, Hugolín: Valaská Škola. Ed. G. J. Sabo. Ohio: Slavica Publishers, Inc. 1987. 730 s.
Pramene k dejinám slovenčiny. Eds. M. Majtán – J. Skladaná. Bratislava: Veda 1992. 397 s.
Pramene k dejinám slovenčiny 2. Eds. T. Laliková – M. Majtán. Bratislava: Veda 2002. 276 s.
Pramene k dejinám slovenčiny 3. Eds. R. Kuchar – I. Valentová. Bratislava: Veda 2008. 296 s.
Urbáre feudálnych panstiev na Slovensku. Zv. 2. Eds. R. Marsina – M. Kušík. Bratislava: Vydavateľstvo SAV 1959. 596 s.
Výhražný list zbojníkov adresovaný mestu Bardejov. 25. 7. 1493.

Prílohy



Obrázok 2: Počet dokumentov podľa storočia a dekády. Plnou čiarou sú vyznačené dekády, prerušovanou storočia. Na zvislej osi je percento dokumentov v danom storočí a promile dokumentov v danej dekáde (t. j. plochy pod oboma čiarami sú porovnateľné).



Obrázok 3: Počet tokenov podľa storočia a dekády. Plnou čiarou sú vyznačené dekády, prerušovanou storočia. Na zvislej osi je percento tokenov v danom storočí a promile tokenov v danej dekáde (t. j. plochy pod oboma čiarami sú porovnateľné).