Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied
Oddelenie Slovenského národného korpusu

ČÍSLO SPRÁVY: SNK JÚĽŠ SAV – 1/2019

SIGNATÚRA: 28157
MDT: 808.54 : 519.765

ARCHÍVNE ČÍSLO SPRÁVY: SNK JÚĽŠ SAV – 1/2019

DRUH SPRÁVY: priebežná za rok 2018

Zmluva č. 0323/2017
o združení prostriedkov na tvorbu a rozvoj Slovenského národného korpusu

Autorka správy: Mária Šimková, PhDr., Ph.D.

Zodpovedná riešiteľka: Mária Šimková, PhDr., Ph.D.

Štatutárny zástupca dodávateľa riešenia úlohy: Gabriela Múcsková, doc. Mgr., PhD.

Kontrolovali: Mgr. Katarína Gajdošová, Ph.D.
Mgr. Ivor Uhliarik

Výtlačok číslo: 1

Stupeň utajenia: 003 nie je predmetom utajenia

Bratislava február 2019

Zmluva č. 0323/2017

Zmluvné strany: Ministerstvo školstva, vedy, výskumu a športu Slovenskej republiky, Ministerstvo kultúry Slovenskej republiky, Slovenská akadémia vied, Jazykovedný ústav Ľ. Štúra SAV

Plnenie úloh pri tvorbe a rozvoji
Slovenského národného korpusu
Jazykovedného ústavu Ľ. Štúra SAV
za rok 2018

Rok 2018 bol druhým rokom riešenia úloh stanovených v Zmluve o združení prostriedkov na tvorbu a rozvoj Slovenského národného korpusu č. 0323/2017 na obdobie 1. 1. 2017 – 31. 12. 2021. V Prílohe č. 2 Zmluvy (Harmonogram riešenia úloh) bolo na tento rok určených 10 hlavných úloh, v jednej úlohe (č. 2) sú obsiahnuté dve rozličné položky. Zároveň sa kontinuálne realizovali viaceré priebežné práce, ktoré sú nevyhnutné na prípravu a sprístupnenie konkrétneho korpusu, novej verzie korpusu, databázy a pod. v danom čase, hoci ich finalizácia a zverejnenie sú v Harmonograme naplánované až na ďalší rok.

Prioritou v roku 2018 bolo dokončenie a vydanie viacerých publikácií rozpracovaných v predchádzajúcom období (úlohy č. 5, 6 a 7 z r. 2017), vytvorenie a sprístupnenie korpusu textov štátnej a verejnej správy a koncepčná práca na terminologických záznamoch z tejto oblasti. Prípravné práce boli naplánované aj v rámci úplne novej úlohy – tvorby korpusu pomenovaných entít.

Doplnenie realizácie úloh z r. 2017

Úloha č. 5: Vydanie Kolokačného slovníka adjektív a publikácie Slovenský národný korpus. Texty, anotácie, vyhľadávania (príručka korpusovej lingvistiky) vo forme tlačených publikácií.

Splnená

Slovník kolokácií prídavných mien v slovenčine autoriek D. Majchráková, K. Chlpíková a K. Bobeková sa dostal na knižné pulty v máji 2018. V januári 2019 mu na pražskej (medzinárodnej) súťaži Slovník roku 2019 bolo udelené **Čestné uznanie poroty**. Publikácia Slovenský národný korpus. Texty, anotácie, vyhľadávania autorského kolektívu M. Šimková, K. Gajdošová, B. Kmeťová, M. Debnár bola vydaná začiatkom júla 2018.

Úloha č. 6: Zabezpečenie vydania Frekvenčného slovníka súčasnej slovenčiny a Retrográdneho slovníka súčasnej slovenčiny v tlačenej podobe.

Splnená

Kolektívne dielo Frekvenčný slovník súčasnej slovenčiny na báze Slovenského národného korpusu (R. Garabík – B. Kmeťová – M. Šimková – M. Zumrík a kol.) vyšlo v januári 2018 a Retrográdny slovník súčasnej slovenčiny – slovné tvary na báze Slovenského národného korpusu (R. Garabík – B. Kmeťová – A. Karčová – K. Bobeková – D. Majchráková – K. Chlpíková), ktorý bol daný do tlače v novembri 2018, vyšiel začiatkom februára 2019.

Úloha č. 7: Dokončenie a vydanie Frekvenčného slovníka hovorenej slovenčiny a monografie o dynamike súčasnej slovenčiny v podobe tlačených publikácií.

Splnená na 95 %

Frekvenčný slovník hovorenej slovenčiny na báze Slovenského hovoreného korpusu pod hlavnou redakciou K. Gajdošovej a M. Šimkovej bol vydaný v októbri 2018. Monografia o dynamike súčasnej slovenčiny autorov M. Šimková, J. Levická, M. Debnár je po posúdení recenzentmi, následných korekciách a redakčných úpravách v štádiu tvorby tlačových podkladov.

Slovník slovných spojení. Podstatné mená, Slovník kolokácií prídavných mien v slovenčine, Frekvenčný slovník súčasnej slovenčiny na báze Slovenského národného korpusu a Frekvenčný slovník hovorenej slovenčiny na báze Slovenského hovoreného korpusu boli ako tituly vydané vydavateľstvom SAV Veda osobitne prezentované na knižnom veľtrhu Bibliotéka v novembri 2018.

Plnenie úloh stanovených na rok 2018

Úloha č. 2: Vytvorenie a sprístupnenie nových verzií minimálne 2 paralelných korpusov.

Splnená 27. 11. a 6. 12. 2018

Spomedzi už existujúcich 9 paralelných slovensko-inojazyčných korpusov SNK boli v r. 2018 vybrané na aktualizáciu 2 korpusy, a to slovensko-latinský paralelný korpus a slovensko-český paralelný korpus.

Slovensko-latinský paralelný korpus (https://korpus.sk/skla.html) patrí medzi rozsahom menšie a svojím obsahom i zameraním raritné korpusy jednak vzhľadom na latinčinu ako mŕtvy jazyk a na jej tri formy, jednak tým, že ide výlučne o preklady z latinčiny do slovenčiny. Zároveň je však tento korpus veľmi dôležitý pri štúdiu klasických jazykov, translatológie, ale aj pri výskumoch v oblasti vývinu filozofického a teologického myslenia či všeobecnej vzdelanosti a funkčnosti slovenčiny ako prekladového jazyka. Predchádzajúca verzia slovensko-latinského paralelného korpusu vytvorená v r. 2014 obsahovala 1,44 mil. tokenov. Nová verzia korpusu **par-skla-3.0** bola sprístupnená 6. 12. 2018 v rozsahu **5 mil.** tokenov (v slovenskej časti 2,66 mil. tokenov, v latinskej časti 2,3 mil. tokenov) a celkovo obsahuje 36 prekladov: 14 z klasickej, 8 zo stredovekej, 14 z novovekej latinčiny.

Slovensko-český paralelný korpus (https://korpus.sk/skcs.html), ktorého predchádzajúca verzia bola vytvorená v r. 2016, bol aktualizovaný v osobitnej časti obsahujúcej beletristické texty. Práve tieto texty sa často používajú na analýzu fungovania jazykových prostriedkov v našich dvoch blízkopríbuzných jazykoch a keďže ich do SNK pribudlo za krátky čas značné množstvo, bolo vhodné aktualizovať tento podkorpus aj v odstupe dvoch rokov. Nová verzia podkorpusu **par-skcs-fic-5.0** bola k 27. 11. 2018 rozšírená o vyše 12 mil. tokenov na rozsah takmer **31,5 mil.** tokenov (v slovenskej časti 15,72 mil. tokenov, v českej časti 15,77 mil. tokenov) a aktuálne obsahuje 217 kníh, z toho 116 preložených zo slovenčiny do češtiny, 56 preložených z češtiny do slovenčiny, 3 napísané jedným autorom v slovenčine aj češtine (V. Zamarovský), 28 textov preložených do slovenčiny aj do češtiny z angličtiny, 14 textov preložených do slovenčiny aj do češtiny z iných jazykov.

Úloha č. 3: Vybudovanie a sprístupnenie prvej verzie paralelného slovensko-poľského korpusu.

Splnená 27. 11. 2018

Najnovší paralelný korpus je 10. korpusom tohto typu v rámci SNK a jeho prvá verzia **par-skpl-1.0** (https://korpus.sk/skpl.html) je k dispozícii v rozsahu takmer **8,2 mil.** tokenov (v slovenskej časti 4,12 mil. tokenov, v poľskej časti 4,06 mil. tokenov). Slovensko-poľský

paralelný korpus obsahuje preklady 42 beletristických textov: 25 z poľštiny do slovenčiny, 6 zo slovenčiny do poľštiny, 11 z iných jazykov do slovenčiny aj poľštiny, ako aj jeden dokument o vzájomnej spolupráci.

Úloha č. 4: Dokončenie a knižné vydanie publikácie *Časovanie slovies v slovenčine s korpusovými príkladmi*.

Dodatok č. 1 k Zmluve: úloha navrhnutá na presunutie do r. 2019

Vzhľadom na množstvo úloh súvisiacich s dokončovaním a vydávaním viacerých publikácií požiadalo riešiteľské pracovisko o presunutie tejto úlohy na rok 2019. Potreba jej presunu sa diskutovala už na prezentačnom dni konanom 16. 5. 2018 v JÚĽŠ SAV, kde sa dohodlo, že v súvise s ďalšími menšími úpravami Zmluvy pripraví JÚĽŠ SAV Dodatok a predloží ho zmluvným partnerom. Obsah Dodatku mal zahŕňať aj zmenu názvu a ďalších náležitostí pracoviska v rámci prebiehajúcej transformácie SAV a jej ústavov. Vzhľadom na problémy vzniknuté pri transformácii sa posunul aj čas prípravy predmetného Dodatku, ktorý bol napokon prediskutovaný a predložený partnerom v októbri 2018. Predsedníctvo SAV s jeho znením súhlasilo bez výhrad, no formálne úpravy legislatívnych pracovníkov ďalších zmluvných partnerov a časovo náročné schvaľovacie procesy spôsobili, že zatiaľ nebol oficiálne odsúhlasený a podpísaný. Návrh na presun úlohy č. 4 do r. 2019 ako jednu z obsahových súčastí Dodatku však nikto zo zmluvných partnerov nespochybnil. Aktuálne sa proces odsúhlasovania Dodatku zrejme zavŕšil na všetkých úrovniach a malo by sa pristúpiť k jeho podpisovaniu.

Úloha č. 8: Spracovanie a sprístupnenie minimálne 1000 nových termínov v Slovenskej terminologickej databáze: kategória *ekonómia*.

Splnená 31. 12. 2018

V r. 2018 sa do Slovenskej terminologickej databázy spracovalo 1 069 nových terminologických záznamov z oblasti **ekonómie**, čím sa celkový počet termínov v STD priblížil k 10 tisícom v 19 kategóriách. Terminologické záznamy z ekonómie (https://terminologickyportal.sk/wiki/Kategória:Ekonómia) boli spracované z viacerých zdrojov poskytnutých do STD na základe licenčnej zmluvy:

Viestová, Kristína a kol.: Lexikón obchodu 1. Trh, obchod, tovar. Bratislava: Vydavateľstvo Ekonóm 2006.
Viestová, Kristína a kol.: Lexikón obchodu 2. Predajňa, obchodný podnik. Bratislava:

Vydavateľstvo Ekonóm 2006.
Viestová, Kristína a kol.: Lexikón logistiky. Bratislava: Vydavateľstvo Iura Edition 2007.

Ak bol v citovanej literatúre uvedený pri definícii iný zdroj, je uvedený ako východiskový zdroj definície aj v STD.

Úloha č. 9: Koncepcia tvorby korpusu textov štátnej správy SR a vytvorenie pilotnej verzie.

Splnená v decembri 2018

Tvorba špecializovaného korpusu textov štátnej správy (na inom mieste v Harmonograme uvedený aj ako korpus textov štátnej a verejnej správy) bola diskutovaná na viacerých pracovných stretnutiach zainteresovaných pracovníkov SNK. V prvom rade sa, aj vzhľadom na plánované využitie týchto textov na tvorbu termínov do Slovenskej terminologickej databázy, vyjasňoval obsah a rozsah termínov *štátna správa* a *verejná správa*. Zároveň sa zvažovalo aj ďalšie začlenenie textov z tejto oblasti do pripravovanej novej verzie hlavného korpusu, kde by sa o niečo rozšírilo doteraz veľmi malé zastúpenie administratívnych textov. Výsledkom analýz a diskusií boli nasledujúce závery a kroky:

- z dostupných internetových zdrojov boli do archívu SNK zaradené dokumenty inštitúcií štátnej a verejnej správy – VÚC, obce, ministerstvá, ostatné orgány štátnej správy v rozsahu 5 440 súborov;
- po konverzii a následnom selektovaní dokumentov s nadštandardným rozsahom tabuliek (viac ako 30 %) bolo 12,67 % súborov vyradených z ďalšieho spracovania;
- pre potreby špecializovaného korpusu textov štátnej správy sa osobitne spracúvali dostupné texty výročných správ ministerstiev a iných štátnych inštitúcií;
- zo spracovaných textov bol vytvorený pilotný korpus ako interná databáza v rozsahu takmer 18 mil. tokenov;
- po dopracovaní a vyplnení príslušnej anotačnej štruktúry v rámci štýlovo-žánrovej anotácie SNK bude korpus sprístupnený v rámci skupiny špecializovaných korpusov a samotné texty budú zaradené aj do novej verzie hlavného korpusu.

Úloha č. 9: Sprístupnenie prvej verzie korpusu textov štátnej správy SR.

Splnená 3. 12. 2018

V rámci úlohy č. 9 bol vytvorený a sprístupnený špecializovaný korpus textov štátnej

správy **gov-web-1.0** (https://korpus.sk/govweb.html) v rozsahu **11,7 mil.** tokenov. Korpus sa skladá z textov štátnych inštitúcií dostupných na webových doménach gov a egov do polovice roka 2017.

Úloha č. 9: Korpusovolingvistické spracovanie textového materiálu z projektu otvorených dát.

Splnená 3. 12. 2018

Dáta dostupné v rámci projektu OpenData obsahujú predovšetkým štatistické, číselné údaje, dajú sa v nich však nájsť aj texty z rôznych oblastí vhodné na korpusové spracovanie. Z časti z nich – zo zdrojov sprístupnených Ministerstvom spravodlivosti SR – bol v SNK vytvorený špecializovaný korpus textov súdnych rozhodnutí **od-justice-1.0** (https://korpus.sk/OpenData.html) v rozsahu vyše **4** mld. tokenov.

Úloha č. 13: Sprístupnenie novej, rozšírenej verzie textového Korpusu nárečí – *dialekt-4.0*.

Splnená 18. 12. 2018

Na príprave textov do novej verzie Korpusu nárečí SNK sa priebežne pracovalo už v r. 2017, keď bola na rôznych úrovniach spracovaná prvá skupina textov (porov. správu za rok 2017, s. 9). V r. 2018 sa celkovo naskenovalo 62 už publikovaných nárečových prepisov, po skenovaní sa rekonštruovalo 33 textov v rozsahu 2 450 normostrán a korigovalo sa 16 textov v rozsahu takmer 763 strán. Do novej verzie z nich bolo spracovaných 28 textových zdrojov z prác viacerých dialektológov, napr. L. Bartka, F. Buffu, A. Ferenčíkovej, Š. Liptáka, I. Ripku. Osobitný príspevok predstavujú krátke nárečové texty publikované v súboroch Zo studnice rodnej reči 1 (Bratislava: VEDA 2005) a Zo studnice rodnej reči 2 (Bratislava: VEDA 2014). V rámci skvalitňovania vnútornej štruktúry Korpusu nárečí SNK došlo k zjednoteniu názvov zdrojov (doc.source) na formát podobný príslušnej položke v anotácii hlavného korpusu.

Aktuálna verzia Korpusu nárečí Slovenského národného korpusu **dialekt-4.0**, ktorý sa v rámci SNK postupne buduje od r. 2013 (https://korpus.sk/dialect.html), bola sprístupnená v rozsahu **711 766** tokenov (verzia 3.0 z r. 2016 obsahovala 494 722 tokenov).

Úloha č. 14: Koncepcia a testovanie ručnej sémantickej analýzy vybranej vzorky textov – anotácie pomenovaných entít a viacslovných spojení.

Splnená v decembri 2018

V rámci riešenia tejto úlohy v SNK bola sémantická analýza textov vymedzená ako anotácia pomenovaných entít. Pomenovaná entita (meno osoby, názov obce, inštitúcie, výrobku atď.) môže byť jednoslovná (*Martin* ako meno osoby i názov mesta) alebo viacslovná (*Slovenská republika*) či vložená v inej pomenovanej entite v ľubovoľnej hĺbke (*Jazykovedný ústav // Ľudovíta Štúra, Jazykovedný ústav // Ľudovíta Štúra // Slovenskej akadémie vied*). Pri anotácii sa uvedeným entitám priraďuje príslušná významová kategória, čo má vplyv na ďalšiu analýzu a spracovanie textu, ale najmä na získavanie informácií z korpusu v širšom, nielen lingvistickom rozsahu.

Po predchádzajúcom štúdiu relevantnej literatúry, dostupných inojazyčných anotácií a korpusov podobného typu a po predbežnej diskusii v SNK, o. i. na dvoch odborných seminároch ešte v r. 2016, sa v priebehu roka 2018 diskutovala koncepcia vybudovania korpusu (otázka výberu textov) a vytvorenia tagsetu (súboru značiek a anotačných pravidiel). Uskutočnilo sa päť pracovných stretnutí, na ktorých sa postupne formovala koncepcia a ktorých výsledkom boli nasledujúce rozhodnutia:

- zdrojom textov na anotáciu budú verejne prístupné dokumenty, predovšetkým články slovenskej Wikipédie, aby výsledný korpus nepodliehal reštriktívnym licenciám;
- dokumenty na anotáciu budú vzorkované rovnomerne podľa kategórií článkov a ucelene (dokumenty budú anotované vo svojom plnom objeme);
- na anotáciu sa vyberie toľko článkov, aby sa celkovo anotovalo aspoň 5 000 viet;
- ručná anotácia sa nebude vykonávať nad čistým dokumentom, ale po predspracovaní, čiže po východiskovej automatizovanej anotácii (predanotácii; porov. ďalej) na základe vopred vytvorených lexikónov;
- hierarchia kategórií pomenovaných entít bude v rámci uniformnosti inšpirovaná Českým korpusom pomenovaných entít (CNEC) a modifikovaná podľa vlastných špecifík;
- korpus pomenovaných entít SNK bude používať rozšírenie XML formátu podobné formátu použitému v CNEC;
- anotátori budú vykonávať prácu pomocou nástrojov, ktoré uľahčia proces ručnej anotácie (porov. ďalej);
- jeden článok bude ručne značkovaný aspoň dvomi anotátormi a následne sa bude

kontrolovať prekrytie výsledkov anotácie;

- anotácie môžu obsahovať IRI (Internationalized Resource Identifier) linky jednoznačne identifikujúce pomenované entity (oblasť entity linking);
- anotovať sa perspektívne môžu aj koreferencie odkazujúce na pomenované entity.

Predanotácia je založená na využití kolekcie lexikónov pomenovaných entít pre identifikované kategórie (zoznam častých slovenských krstných mien, obcí, televíznych staníc, krajín atď.), ktoré boli ručne vyčistené a dezambiguované podľa relevantných častí morfologických značiek z tagsetu SNK. Výsledná kolekcia lexikónov obsahuje 228 950 kategorizovaných pomenovaných entít v 19 kategóriách. Na predanotáciu dokumentov na základe týchto lexikónov, morfologickej analýzy textu a blízkej lematizácie bol navrhnutý a implementovaný príslušný algoritmus.

Na uľahčenie práce ručnej anotácie boli vyvinuté nástroje vo forme rozšírenia pre editor vim, resp. jeho nadstavby (napr. grafický gvim alebo cream). Pomocou nich je farebne zvýraznená syntax navrhnutého formátu, rozlišujú sa kategórie označených entít a anotátori majú k dispozícii klávesové skratky na určovanie a úpravu anotácií. Predanotácia a nástroj pre anotáciu boli testované na vybraných článkoch Wikipédie.

Dlhodobejším cieľom je vybudovanie **Slovenského korpusu pomenovaných entít** – korpusu slovenských textov s ručnou anotáciou pomenovaných entít (v štandardnom zmysle v oblasti spracovania prirodzeného jazyka), ktorý zatiaľ pre slovenčinu v takomto rozsahu neexistuje. Na splnenie tohto cieľa sú naplánované čiastkové úlohy aj v ďalších rokoch projektu.

Úloha č. 15: Pracovné semináre o používaní korpusových zdrojov podľa požiadaviek záujemcov.

Plnená priebežne

Praktické semináre pre začínajúcich aj pokročilých používateľov zdrojov SNK sú stále žiadané najmä vo vysokoškolskom prostredí, kde sú mnohé postupové práce založené na textových a jazykových databázach SNK, ale aj medzi inými záujemcami (zahraniční študenti slovenčiny, prekladatelia, učitelia nižších stupňov škôl). Tieto semináre sa uskutočňujú podľa požiadaviek záujemcov (i podľa možností pracovníkov SNK) priebežne počas celého roka. V r. 2018 ich bolo spolu 15 (v Bratislave, Banskej Bystrici, Revúcej a Košiciach) a zúčastnilo sa na nich celkovo 227 študentov, pedagógov, vedecko-výskumných pracovníkov a iných

záujemcov.

Okrem vyššie uvedených výsledkov boli navyše oproti Harmonogramu riešených úloh vytvorené a verejnosti sprístupnené ďalšie tri zdroje:

- 19. 2. 2018 – referenčný korpus **prim-7.0-frk** (https://korpus.sk/ref.html) v rozsahu **253 137 609** tokenov, ktorý bol vytvorený z hlavného korpusu prim-7.0-public-all na základe štyroch hlavných kritérií vychádzajúcich z koncepcie Frekvenčného slovníka slovenčiny na báze Slovenského národného korpusu; z korpusu prim-7.0-frk boli napočítané hodnoty aj pre Retrográdny slovník súčasnej slovenčiny – slovné tvary na báze Slovenského národného korpusu;
- 2. 5. 2018 – piata verzia korpusu textov z Wikipédie a Necyklopédie **wiki-2018-03** (https://korpus.sk/wiki.html) v rozsahu **47 283 205** tokenov;
- 20. 12. 2018 – tretia verzia databázy Archívu nárečí SNK rozšírená o informácie o **53** digitalizovaných nárečových nahrávkach (https://korpus.sk/dialect_recordings.html).

V plnom rozsahu sa v r. 2018 realizovali aj nasledujúce priebežné úlohy stanovené v Harmonograme riešenia úloh.

Úloha č. 1: Sťahovanie a spracúvanie textov povinných výtlačkov.

Realizovalo sa priebežne podľa interného harmonogramu SNK.

Úlohy č. 1, 2, 10, 11 – 13: Dopĺňanie a) všeobecného korpusu, b) paralelných korpusov a c) ďalších korpusov SNK aktuálnymi textami.

Naskenovalo sa približne 30 500 strán textov z knižných a časopiseckých titulov, rekonštruovalo, opravilo a skontrolovalo sa približne 22 000 strán textov. Do archívu SNK bolo vložených 938 nových dokumentov s príslušnou anotáciou.

Úlohy č. 1, 2, 10, 11 – 13: a) Oslovovanie poskytovateľov textov, b) spracúvanie licenčných zmlúv, c) správa databázy poskytnutých textov.

Nové licenčné zmluvy boli uzavreté s 10 poskytovateľmi (napr. s TASR). Aj v r. 2018

sa však tak ako v predchádzajúcom roku dopĺňali do jednotlivých častí SNK predovšetkým tie texty, na ktoré boli uzavreté licenčné zmluvy v minulosti, kontrolovala sa korešpondencia záznamov v archíve a databáze, aktualizovali sa kontakty s poskytovateľmi, prebiehala komunikácia s poskytovateľmi povinných výtlačkov.

Úlohy č. 1, 2, 10, 11 – 13: Konverzie textov do jednotného formátu.

Z dokumentov vložených do archívu SNK bolo skonvertovaných 662 dokumentov, viaceré konverzné skripty bolo potrebné upraviť.

Úlohy č. 1, 2, 10, 11 – 13: Bibliografická a štýlovo-žánrová anotácia textov.

Do banky SNK pribudlo 394 nových bibliografických a štýlovo-žánrových záznamov.

Úloha č. 15: a) Dopĺňanie morfologického slovníka, b) skvalitňovanie anotácií, c) anotačných a vyhľadávacích nástrojov.

Po predchádzajúcich rozsiahlejších zjednocovaniach v morfologickej databáze sa v tomto roku priebežne realizovalo len niekoľko drobných zjednotení a opráv.

Úlohy č. 8, 10: a) Zhromažďovanie a b) spracúvanie odborných textov pre databázu termínov vybraných vedných odborov, c) dopĺňanie a aktualizácia Slovenskej terminologickej databázy.

Body a) a b) sa napĺňali predovšetkým spracúvaním textov verejnej a štátnej správy (porov. vyššie Úloha č. 9). Najrozsiahlejšie doplnenie Slovenskej terminologickej databázy sa uskutočnilo v rámci Úlohy č. 8 (1 000 nových termínov z oblasti ekonómie), ale rozpracovaných bolo ďalších 400 nových terminologických záznamov z oblasti IT a gastronómie. Pokračovalo sa aj v rozširovaní a skvalitňovaní terminologického portálu, do ktorého sa dopĺňali nové terminologické zdroje s anotáciami.

Úloha č. 11: a) Zhromažďovanie a b) spracúvanie zvukových záznamov do hovoreného korpusu.

V rámci prípravy novej verzie Slovenského hovoreného korpusu, ktorej finalizácia a sprístupnenie sú plánované na r. 2020, bolo v r. 2018 prepísaných 15 nahrávok v rozsahu takmer 20 hodín a skorigovaných 36 nahrávok v rozsahu 44 hod. 25 min.

Úloha č. 12: a) Získavanie a b) spracúvanie pramenných materiálov do historického korpusu.

Do novej verzie Historického korpusu slovenčiny, ktorá je naplánovaná na rok 2019, sa na základe konzultácií s pracovníkmi Oddelenia dejín slovenčiny, onomastiky a etymológie JÚĽŠ SAV vyberali na spracovanie texty z archívu tohto oddelenia. Prepis pôvodných textov receptárov (rád a pod.) realizujú dve anotátorky.

Úloha č. 13: a) Zhromažďovanie a b) spracúvanie zvukových záznamov do nárečového korpusu.

Priebežne budovaný Archív nárečí SNK bol rozšírený o 109 nových nahrávok, z nich bolo 55 nahrávok technicky upravovaných – čiastočne zbavených šumu. Jedna nahrávka v rozsahu 49 minút 24 sekúnd bola transkribovaná.

Úloha č. 15: Zabezpečovanie korpusov a databáz efektívnymi a aktuálnymi nástrojmi vhodnými na lingvistické využitie i počítačové spracovanie prirodzeného jazyka.

V tejto úlohe sa realizovali iba najnevyhnutnejšie práce vzhľadom na nedostatočné personálne obsadenie v skupine IT pracovníkov.

Úloha č. 15: Prezentácie materiálových a textových zdrojov SNK odbornej i laickej verejnosti.

Okrem pracovných seminárov uvedených v úlohe č. 15 pracovníci SNK pravidelne prezentujú pracovisko na podujatí Európska noc výskumníkov, v rámci Dňa otvorených dverí v JÚĽŠ SAV, uskutočňujú prednášky o korpusových databázach pre prekladateľov (napr. na FF UKF v Nitre), ako aj pre bežných záujemcov (napr. pre gymnazistov z Topoľčian, študentov slovenčiny z Krakova). Tento rok prezentovali aj viaceré knižné publikácie na knižnom veľtrhu Bibliotéka a zúčastnili sa podujatia Víkend so SAV.

osobitné Terminologická poradenská služba.

Realizuje sa v terminologických komisiách, kde má SNK JÚĽŠ SAV zastúpenie, ako aj formou mailov či priebežnými osobnými konzultáciami.

osobitné Korpusovolingvistická poradenská služba.

Bolo poskytnutých 25 osobných a telefonických konzultácií, ďalšie otázky boli zodpovedané pri prideľovaní nových alebo obnovovaných registráciách na používanie SNK.

V r. 2018 bolo celkovo registrovaných 590 aktívnych používateľov SNK z 15 rôznych krajín.

osobitné Správa a aktualizácia a) webovej stránky SNK a b) prezentačnej stránky SNK na Facebooku.

Informácie o nových zdrojoch SNK a ďalšie aktuality sú zverejňované v novinkách na stránke https://korpus.juls.savba.sk a na Facebooku, kde má SNK vyše 1 110 sledovateľov.

osobitné Správa a aktualizácia a) počítačovej siete, b) pracovných staníc, c) úložísk, d) záložných zdrojov.

Realizovala sa priebežná údržba, oprava, montáž, výmena a aktualizácia počítačov (fyzických a virtuálnych) a iných hardvérových i softvérových nástrojov. Na aktualizáciu počítačového vybavenia a zlepšenie infraštruktúry SNK (zabezpečenie bezporuchového využívania dát externými používateľmi, zálohovanie dát a pod.) bol po analýze ponúk zostáv servera a finančných možností pracoviska zakúpený nový server.

Vypracovala: PhDr. Mária Šimková, Ph.D.
zodpovedná riešiteľka