Plnenie úloh pri tvorbe a rozvoji Slovenského národného korpusu za rok 2019

V Zmluve o združení prostriedkov na tvorbu a rozvoj Slovenského národného korpusu č. 0323/2017 sa ako predmet na celú dobu riešenia (1. 1. 2017 – 31. 12. 2021) špecifikuje celkovo 16 položiek, ktoré zahŕňajú vyše 50 úloh s ďalšími nadväzujúcimi priebežnými prácami. Po troch rokoch realizácie tejto etapy budovania Slovenského národného korpusu a plnenia súvisiacich úloh môžeme konštatovať splnenie 34 čiastkových úloh a v rámci nich úplné uzavretie 6 položiek obsahujúcich o. i. dokončenie a vydanie viacerých slovníkov a monografií. Príprava väčšiny doterajších publikačných výstupov si síce vyžiadala viac času a práce, ako bolo pôvodne plánované, ale priniesla svoje výsledky aj v podobe ocenení ich kvality. Po Čestnom uznaní poroty pražskej (medzinárodnej) súťaže Slovník roku 2019, ktoré bolo minulý rok udelené dielu *Slovník kolokácií prídavných mien v slovenčine* (autorky D. Majchráková, K. Chlpíková, K. Bobeková), sa na tej istej súťaži tento rok zúčastnil *Retrográdny slovník súčasnej slovenčiny – slovné tvary na báze Slovenského národného korpusu* (autori R. Garabík, B. Kmeťová, A. Karčová, K. Bobeková, D. Majchráková, K. Chlpíková), ktorý získal 2. miesto v súťaži o hlavnú cenu Slovník roku 2020 a Cenu poroty. Ten istý slovník bol ocenenený aj Prémiou Literárneho fondu za vedeckú a odbornú literatúru za r. 2018. Hlavná cena Literárneho fondu za vedeckú a odbornú literatúru za rok 2018 bola udelená titulu *Frekvenčný slovník hovorenej slovenčiny na báze Slovenského hovoreného korpusu* (autori K. Gajdošová, M. Šimková a kolektív).

Na rok 2019 bolo v Harmonograme riešenia úloh (Príloha č. 2 Zmluvy) stanovených 11 položiek, pričom vo viacerých prípadoch išlo o čiastkové zavŕšenie priebežných prác.

Úloha č. 1: Sprístupnenie novej, skvalitnenej a rozšírenej verzie Slovenského národného korpusu – hlavného korpusu prim-9.0 a jeho podkorpusov.

Splnená vo februári 2020

Práce na príprave, anotáciách a dopĺňaní textov do novej verzie hlavného korpusu sa zavŕšili koncom roka 2019, keď zároveň prebehli aj prvé procesy finalizácie a plošného

zjednocovania spracovania textov. Začiatkom roka bola vytvorená kompletná nová verzia, po jej otestovaní sa však ukázalo, že niektoré procesy výsledného spracovania neprebehli korektne a do korpusu sa nedostalo značné množstvo textov pripravených v priebehu r. 2019. Pri hľadaní a riešení vzniknutého problému sa odhalili ďalšie technické nedostatky, ktorých odstraňovanie oddialilo definitívne sprístupnenie novej verzie korpusu. V záujme zahrnutia všetkých spracovaných textov (najmä aktuálnej lexiky) v čo najlepšej kvalite sa zvážilo, že bude vhodné investovať do finalizácie tejto verzie ďalšiu prácu, aj keď sa posunul termín sprístupnenia nového korpusu všetkým používateľom.

Aktuálne spracované dáta (texty ku koncu roka 2019) po lematizácii a morfologickej anotácii, ktoré už sú interne k dispozícii, predstavujú v nefiltrovanom objeme rozsah vyše 2 mld. tokenov, čo je nárast oproti predchádzajúcej verzii o takmer 270 mil. tokenov. Korpus prim-9.0-juls-all, ktorý bude prístupný v rámci JÚĽŠ SAV, bude po odfiltrovaní v rozsahu necelých 2 mld. tokenov. Najnovšia verzia hlavného korpusu prim-9.0-public-all bude z toho obsahovať verejne prístupné texty SNK v rozsahu približne 1,7 mld. tokenov. Rozdiel medzi interným a verejne sprístupneným korpusom predstavujú o. i. texty bez súhlasu na ich zverejnenie v SNK, no dôležité na lingvistický výskum (napr. staršie časopisy Kultúrny život, Pyramída). O možnosti zaradenia takýchto textov do verejne prístupných databáz sme opakovane rokovali s viacerými znalcami v oblasti autorských práv, zatiaľ bez výsledku.

Okrem kvantitatívneho nárastu dát v hlavnom korpusovom zdroji SNK sa do novej verzie v r. 2018 a 2019 priebežne skvalitňovali mnohé vstupné texty v rámci korekcií konverzných procesov, zlepšenia selekcie cudzojazyčných častí textov a pod., odhaľovali a opravovali sa niektoré typy chýb, korigovali a zjednocovali sa anotačné záznamy bibliografickej a štýlovo-žánrovej anotácie. Korpus aktuálne tvorí vyše 1,8 mil. dokumentov, napr. v kategórii autor je približne 35 tisíc položiek, v kategórii bibliografia takmer 50 tisíc položiek. Keďže ide prevažne o manuálne vypĺňané záznamy, na ktorých sa podieľajú viacerí pracovníci, objavujú sa v nich drobné chyby a nejednotnosti. Niektoré kľúče a hodnoty v anotácii bolo potrebné priebežne meniť alebo dopĺňať podľa aktuálne pribúdajúcich textov a ich vlastností, čo malo vplyv na staršie zaradenia podobných textov a ich anotácie. Komplexná revízia anotačných záznamov je naplánovaná na rok 2020.

Úloha č. 2: Sprístupnenie novej, skvalitnenej a rozšírenej verzie webového korpusu slovenčiny – web-5.0.

Splnená 27. 1. 2020

Webový korpus web-5.0 bol sprístupnený v rozsahu vyše 4 mld. tokenov, čo je nárast oproti predchádzajúcej verzii o vyše 1 mld. tokenov. Dáta do tejto verzie webového korpusu boli poskytnuté z projektu Aranea v rámci spolupráce medzi oddeleniami JÚĽŠ SAV. Pri príprave a prvotnom spracovaní dát sa uplatnili najnovšie postupy a skúsenosti z tvorby tohto typu korpusu. Následne boli tieto dáta v SNK tokenizované, segmentované, lematizované a morfologicky označkované štruktúrou a značkami v zásade kompatibilnými s písanými korpusmi SNK.

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

Splnená 17. 7. 2019

Prvá verzia slovensko-španielskeho paralelného korpusu bola sprístupnená 17. 7. 2019 v rozsahu takmer 11,5 mil. tokenov (5 455 067 tokenov v slovenskej časti, 6 044 520 tokenov v španielskej časti). Korpus obsahuje v tejto verzii preklady 77 textov, z toho 59 zo španielčiny do slovenčiny, 1 zo slovenčiny do španielčiny a 17 z iných jazykov do slovenčiny aj španielčiny. Texty sú štandardne zarovnané po vetách, lematizované a morfologicky anotované v obidvoch jazykoch.

Úloha č. 4: Koncepcia a príprava publikácie *Tvary prídavných mien s korpusovými príkladmi*.

Splnená v októbri 2019

Od začiatku roka sa k tejto téme realizovalo viacero pracovných stretnutí a konzultácií, opravoval a zjednocoval sa východiskový materiál (najmä vo vzťahu k homonymným príčastiam). Koncepcia a dosiahnutý stav prác na príprave opisu tvarov prídavných mien boli prezentované a diskutované na internom seminári SNK 21. 10. 2019. Výsledkom sú rozpracované úvodné texty (cca 13 strán), v ktorých sú stručne charakterizované vlastnosti a triedenie prídavných mien v slovenčine a opísaná koncepcia publikácie. Tak ako v prípade už publikovaného prehľadu skloňovania podstatných mien na báze SNK (2016), aj tu sa vychádza z paradigiem prídavných mien spracovaných v morfologickej databáze Slovenského národného korpusu a rozdelených podľa formálneho princípu (Levensteinove edit-operácie). Z hľadiska skloňovania sa v tejto fáze ukazuje, že

prídavné mená možno rozdeliť do 33 skupín paradigiem, z hľadiska stupňovania do ďalších 20 skupín paradigiem. Reprezentanty príslušného skloňovania s plnými paradigmami a zoznamami slov s rovnakým skloňovaním a stupňovaním sú súčasťou pripraveného materiálu v samostatných súboroch.

Úloha č. 8: Spracovanie a sprístupnenie minimálne 300 nových termínov v Slovenskej terminologickej databáze: kategória *štátna správa*.

Splnená 31. 12. 2019

Na základe analýzy relevantných zdrojov sa ukázalo, že samotné pomenovanie kategórie je terminologicky vhodnejšie v podobe *verejná správa*. V tejto novej kategórii bolo vytvorených a sprístupnených 336 nových terminologických záznamov.

Okrem tohto plánovaného rozšírenia Slovenskej terminologickej databázy sa vytvoril priestor na doplnenie a revidovanie terminologických záznamov z oblasti medicíny (1 900 nových záznamov), na aktualizáciu a sprístupnenie terminologických záznamov z oblasti masmediálnej komunikácie (580 záznamov) a na niekoľko ďalších menších úprav a doplnení. Slovenská terminologická databáza aktuálne obsahuje vyše 12 tisíc terminologických záznamov z 21 oblastí.

Úloha č. 12: Vytvorenie a sprístupnenie novej verzie Historického korpusu slovenčiny – *hist-5.0*.

Splnená 24. 2. 2020

V r. 2019 sa završovali práce začaté v predchádzajúcom roku na vlastnom prepise 5 pôvodných historických textov (Poznámky a rady hygienické, prírodopisné, meteorologické, ekonomické, Receptár gemerský, Receptár nitriansky, Receptár osturniansky – zápisy o ľudovom liečení, Receptár prešovský). Prepisy sa korigovali, zjednocovali a dopĺňali o relevantné informácie v štruktúrnych značkách a komentároch. Výsledkom je nová verzia korpusu v rozsahu takmer 1 mil. tokenov, nárast oproti predchádzajúcej verzii predstavuje vyše 80 tisíc tokenov. Historický korpus slovenčiny spolu s Korpusom nárečí SNK a Archívom nárečí SNK boli prezentované na vedeckom seminári Jazyková zmena a možnosti jej skúmania 17. 10. 2019 na pôde FF UK v Bratislave.

Úloha č. 14: Ručná sémantická analýza vybranej vzorky textov – anotácia pomenúvajúcich entít a viacslovných spojení.

Plnená priebežne

Na základe koncepcie a plánu prác, ktoré boli vypracované v r. 2018, sa pristúpilo k ručnej anotácii pomenovaných entít. Na anotáciu sa vyberali texty z korpusu slovenskej Wikipédie (porov. ďalej) z hľadiska obsahu (rôzne témy podľa možnosti v rovnomernom zastúpení) a rozsahu (objem textu). V rámci testovacej fázy bolo anotovaných 16 článkov, ktoré tvorili 54 súborov v rozsahu vyše 10 tisíc viet. 15 článkov bolo anotovaných aspoň dvoma anotátorkami, 1 článok bol anotovaný 1 anotátorkou. Celkovo bolo označených vyše 20 tisíc entít, ktoré však v tomto testovacom súbore nemusia úplne zodpovedať výsledným pravidlám anotácie, ale môžu odrážať priebežné zmeny a úpravy po testovaní korektnosti nastavených pravidiel.

V ďalšej fáze anotácie, realizovanej už na základe ustálených pravidiel, bol vytvorený súbor dát pre plánovanú pilotnú verziu korpusu pomenovaných entít. Tento súbor je zložený zo 43 článkov (spolu 69 súborov) v rozsahu 16 618 viet, v ktorých bolo označených vyše 40 tisíc entít. 26 článkov bolo anotovaných aspoň dvomi anotátorkami.

Úloha č. 14: Tvorba nástrojov na automatizované rozpoznávanie a anotáciu pomenúvajúcich entít a viacslovných spojení.

Plnená priebežne

V procese vyššie opísanej ručnej anotácie sa používal aj nástroj na automatizovanú predanotáciu pomenovaných entít, pomocou ktorého boli entity vopred rozpoznané a vyznačené a anotátorky ich následne manuálne korigovali. S týmto nástrojom sa začínalo pracovať už v r. 2018, v priebehu aktuálnej anotácie v r. 2019 bol vylepšený rozšírením o stavový automat na presnejšiu klasifikáciu pomenovaných entít a rozpoznanie niektorých vložených entít. Zlepšenie sa dosiahlo aj vďaka korekcii kategorizovaného lexikónu pomenovaných entít, ktorý bol vytvorený v predchádzajúcom období.

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

Plnená priebežne

Na 4 prezentáciách pre rôznych poslucháčov a na 8 praktických seminároch pre začínajúcich a pokročilých používateľov SNK sa zúčastnilo celkovo takmer 350 záujemcov z vysokoškolských pracovísk z Košíc, Banskej Bystrice, Trnavy, Bratislavy, zahraniční študenti i lektori Letnej školy Studia Academica Slovaca, študenti strednej školy.

Úloha č. 16: Príprava lexikálno-gramatického opisu adjektív v písanej a hovorenej slovenčine na báze korpusového materiálu – pilotné štúdie.

Plnená priebežne

Na začiatku roka 2019 sa realizovali konzultácie s viacerými spoluriešiteľmi a predbežne sa rozdelili témy na spracovanie do pilotných štúdií. Do výberu sa dostali najmä také typy a javy súvisiace s prídavnými menami, ktoré neboli doteraz v centre pozornosti, napr. deminutívne a augmentatívne tvorenie prídavných mien, opis takto utvorených prídavných mien v slovníkoch a ich fungovanie v komunikácii. O týchto témach boli napísané a do tlače odovzdané 2 štúdie, v ďalšej publikovanej štúdii bola predstavená spájateľnosť a pragmatické aspekty fungovania jedného prídavného mena (*politický*) z hľadiska zmien v spoločnosti a dynamiky jazyka. Záber sa v druhom polroku 2019 rozšíril na analýzu synonymných vzťahov vo vybranej skupine adjektív, na využitie prídavných mien v konkrétnych literárnych dielach, na fungovanie nízko frekventovaných adjektívnych podôb príčastí v textoch z rôznych štýlov a vecných oblastí. Tieto témy boli predstavené na konferencii v 3 prezentáciách, z ktorých boli publikované abstrakty. Príslušné štúdie sa dokončovali na prelome rokov a vo februári 2020 boli odovzdané do tlače.

Úloha č. 15: Organizácia 10. ročníka medzinárodnej konferencie Slovko a publikovanie príspevkov.

Splnená 23. – 25. 10. 2019

Medzinárodná konferencia SLOVKO 2019. Počítačové spracovanie jazyka, korpusová lingvistika, zmeny a dynamika jazyka (SLOVKO 2019. NLP, Corpus Linguistics, Language Dynamics and Change) sa konala 23. – 25. októbra 2019 v Hoteli Devín v Bratislave. Jubilejný 10. ročník tejto konferencie bol zameraný predovšetkým na prezentáciu výsledkov výskumov v oblasti počítačového spracovania prirodzeného jazyka, korpusovej lingvistiky a dynamiky jazyka, osobitne na zmeny v súčasných jazykoch na báze korpusového materiálu. Zastúpené boli aj oblasti e-lexikografie, e-terminológie, analýzy a opisu jazykových zdrojov

pre menej spracované jazyky, automatizovaného prekladu, nástrojov textovej analýzy, tvorby a využitia jazykových zdrojov, hovorených a nárečových korpusov, rečovej analýzy a syntézy. Na konferencii bolo prítomných celkovo 77 účastníkov, z toho 33 bolo z rôznych pracovísk SR a 44 z Bulharska, Českej republiky, Francúzska, Holandska, Nemecka, Poľska, Rakúska, Ruska, Švajčiarska, Švédska a Ukrajiny. Príspevky z konferencie, medzi ktorými malo zastúpenie aj pracovisko SNK, boli publikované v samostatnom tematickom čísle Jazykovedného časopisu, ktoré bolo v plnom rozsahu v anglickom jazyku a ktoré pripravili editori z kolektívu SNK.

Viaceré prezentované výsledky boli založené na priebežnom plnení úloh, z ktorých boli v r. 2019 najdôležitejšie tieto:
Úloha č. 1: Sťahovanie a spracúvanie textov povinných výtlačkov.
Úlohy č. 1, 2, 10, 11 – 13: Dopĺňanie a) všeobecného korpusu, b) paralelných korpusov a c) ďalších korpusov SNK aktuálnymi textami.
Úlohy č. 1, 2, 10, 11 – 13: a) Oslovovanie poskytovateľov textov, b) spracúvanie licenčných zmlúv, c) správa databázy poskytnutých textov.
Úlohy č. 1, 2, 10, 11 – 13: Konverzie textov do jednotného formátu.
Úlohy č. 1, 2, 10, 11 – 13: Bibliografická a štýlovo-žánrová anotácia textov.
Úlohy č. 8, 10: ... c) dopĺňanie a aktualizácia Slovenskej terminologickej databázy.
Úloha č. 12: a) Získavanie a b) spracúvanie pramenných materiálov do historického korpusu.
Úloha č. 15: Prezentácie materiálových a textových zdrojov SNK odbornej i laickej verejnosti.

Nové texty sa získavali jednak oslovovaním nových poskytovateľov textov, jednak sa opäť oslovovali už zazmluvnení poskytovatelia, od ktorých texty neboli dodané alebo medzitým vydali nové texty. Z knižných predlôh, neexistujúcich alebo nedostupných v elektronickej podobe, sa digitalizovalo takmer 20 000 strán textov, zrekonštruovalo, opravilo a skontrolovalo sa vyše 22 000 naskenovaných strán. Obnovovali sa existujúce a uzatvárali sa nové licenčné zmluvy na poskytnutie textov, rokovalo sa o aktuálnych právnych úpravách a doplneniach či revízii niektorých paragrafov v licenčných zmluvách. Aktualizovali sa viaceré publicistické texty od poslednej verzie korpusu a získavali sa najmä odborné texty. Do archívu SNK bolo vložených 729 nových dokumentov, realizovali sa konverzie nových textov (798 jednotlivých a 94 538 automatických konvertov). Upravovali sa

konverzné skripty, opravovali a zjednocovali sa konverzie a anotácie starších textov (1 131 jednotlivých a 143 689 automatických konvertov). Do banky SNK pribudlo 1 010 nových štýlovo-žánrových anotácií textov, opravených bolo 262 starších anotácií. Systematické spracúvanie zdrojov – nahrávok a ich prepisov sa realizovalo v podprojekte Slovenského hovoreného korpusu, kde bolo prepísaných 42 nahrávok v rozsahu vyše 58 hod. a skorigovaných 27 nahrávok v rozsahu vyše 40 hod. Do Historického korpusu slovenčiny bolo prepísaných 5 nových textov v rozsahu vyše 250 strán.

Okrem vyššie uvedených zdrojov bola v priebehu roka 2019 vytvorená a sprístupnená nová verzia korpusu slovenských textov z Wikipédie wiki-2019-08 v rozsahu 50,1 mil. tokenov. Nová verzia tohto korpusu bola značne prepracovaná aj v súvislosti s potrebami zdrojových textov pre vytváraný korpus pomenovaných entít. Oproti predchádzajúcim verziám sa okrem pridania nových textov urobilo viacero dôležitých zmien a opráv: opravy chýb spracovania MediaWiki značiek v predchádzajúcich verziách; matematické výrazy sa spracovali ako jeden token; pridala sa štruktúra <g/> na označenie miest medzi tokenmi, kde sa pôvodne nenachádzala medzera; do tejto verzie sa nezaradili texty Necyklopédie.

Registrovaných aktívnych používateľov korpusových zdrojov SNK bolo na konci r. 2019 celkovo 727. Pracovníci SNK prezentovali pracovisko na Vedeckom veľtrhu, na medzinárodnom festivale vedy Európska noc výskumníkov 2019 a počas Dňa otvorených dverí JÚĽŠ SAV. Na facebookovej stránke SNK bolo ku koncu roka aktuálne 1 255 záujemcov.

Úlohy stanovené v zmluve na rok 2019 boli v plnom rozsahu splnené, v niektorých bodoch aj prekročené (rozšírenie Slovenskej terminologickej databázy, korpus textov z Wikipédie). Osobitná pozornosť sa venovala dokončeniu a vydaniu publikácií, ktoré boli pri predchádzajúcom hodnotení tesne pred dokončením (Šimková, M. – Levická, J. – Debnár, M.: *Dynamické javy v súčasnej slovenčine a jej výskume.* Bratislava: Veda 2018. 196 s.) alebo boli súčasťou Dodatku č. 1 k zmluve (*Človek a jeho jazyk 4. Terminologické inšpirácie profesora Jána Horeckého/Man and His Language 4. Selected Terminological Papers of J. Horecký.* Red. J. Levická – M. Zumrík. Bratislava: Veda 2019. 400 s.). Na finalizácii ďalších dvoch titulov, ktoré mali byť podľa Dodatku č. 1 vydané v r. 2019 (Časovanie slovies v slovenčine a Slovenský národný korpus. Používanie, príklady, postupy), sa takisto intenzívne pracovalo a v tomto čase sa už pripravujú na vydanie.

Vypracovala: PhDr. Mária Šimková, Ph.D.
vedúca projektu
SNK JÚĽŠ SAV