Plnenie úloh Zmluvy o združení prostriedkov č. 0323/2017 podľa špecifikácie predmetu zmluvy (príloha č. 1) v rokoch 2017 – 2021

Do tohto prehľadu prác oddelenia Slovenského národného korpusu boli zaradené hlavné výstupy za jednotlivé roky podľa špecifikácie predmetu Zmluvy. Podrobnejšie informácie ku každému z nich možno nájsť jednak na webovej stránke SNK a jednak v súhrnných správach za príslušné roky.

1. Rozširovanie všeobecného korpusu a špecifických podkorpusov v súlade s dynamikou jazyka a aktuálnymi potrebami jeho výskumu a počítačového spracovania.

SPLNENÁ

Verzia hlavného korpusu	Rozsah verzie	Sprístupnenie verzie
prim-8.0 a jeho podkorpusy	1 477 miliónov tokenov	január 2018
prim-9.0 a jeho podkorpusy	1 652 miliónov tokenov	február 2020
prim-10.0 a jeho podkorpusy	zatiaľ neurčený	marec 2022

Vo verzii **prim-8.0** bola skvalitnená morfologická anotácia a lematizácia (napr. sa začalo rozlišovanie lem s veľkým a malým písmenom). Pri verzii **prim-9.0** sa z hľadiska kvality zlepšili mnohé vstupné texty v rámci korekcií konverzných procesov, selekcie cudzojazyčných častí textov a pod., odhaľovali a opravovali sa niektoré typy chýb, korigovali a zjednocovali sa anotačné záznamy bibliografickej a štýlovo-žánrovej anotácie. Verzia **prim-10.0** takisto predstavuje kvalitatívny skok v podobe prechodu na dokonalejšie softvérové nástroje, vylepšenia konfigurácie a zavedenia nástrojov a modelov vytvorených na pôde SNK ako aj na iných pracoviskách JÚĽŠ SAV, čo sa premietlo do kvalitnejšieho spracovania textov, menovite ich segmentácie, tokenizácie a lematizácie. K vylepšeniu morfologickej anotácie prispelo aj rozšírenie morfologickej databázy.

2. Vytváranie a sprístupňovanie nových verzií webového korpusu slovenčiny a existujúcich paralelných korpusov.

SPLNENÁ

A. tvorba a sprístupňovanie nových verzií webového korpusu slovenčiny:

Verzia webového korpusu	Rozsah verzie	Sprístupnenie verzie
webový korpus web-4.0	2 963 miliónov tokenov	31. 1. 2018
webový korpus web-5.0	4 042 miliónov tokenov	27. 1. 2020
webový korpus web-6.0	4 390 miliónov tokenov	marec 2022

Z hľadiska kvantity bol objem dát vo webovom korpuse v sledovanom období takmer zdvojnásobený (webový korpus web-3.0, sprístupnený 6. 3. 2015, mal rozsah 2 372 miliónov tokenov). Zároveň bolo dosiahnuté obdobné kvalitatívne zlepšenie morfologickej anotácie a lematizácie ako je uvedené v bode 1 pri verziách hlavného korpusu.

B. tvorba a sprístupňovanie nových verzií existujúcich paralelných korpusov:

1) Nová verzia slovensko-latinského paralelného korpusu **par-skla-3.0** bola verejnosti sprístupnená 6. 12. 2018 v rozsahu 5 miliónov tokenov. Nachádza sa v nej 36 prekladov z latinčiny (14 z klasickej, 8 zo stredovekej, 14 z novovekej latinčiny), pričom dva texty sú preklady z pôvodne talianskeho a kombinovaného textu. Texty sú automaticky zarovnané po vetách. Slovenské texty sú automaticky morfologicky anotované tagerom MorphoDiTa natrénovaným v SNK na báze tagsetu vypracovaného v Slovenskom národnom korpuse, latinské texty sú anotované TreeTaggerom.

2) Nová verzia slovensko-českého paralelného korpusu **par-skcs-fic-5.0** bola zaradená medzi verejné korpusové databázy 27. 11. 2018 v rozsahu 31,5 miliónov tokenov. Bolo do nej zahrnutých 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. Všetky texty sú automatizovane zarovnané po vetách. Slovenské texty sú morfologicky anotované tagermi Morče a MorphoDiTa natrénovanými v SNK na báze tagsetu vypracovaného v Slovenskom národnom korpuse, české texty sú anotované tagerom Morče a MorphoDiTa na báze tagsetu použitého v Českom národnom korpuse.

3. Vytvorenie slovensko-poľského a slovensko-španielskeho paralelného korpusu pre potreby porovnávacích výskumov, výučby jazykov a prípadnú tvorbu nových prekladových slovníkov.

SPLNENÁ

A. **Slovensko poľský paralelný korpus** je 10. korpusom tohto typu v rámci SNK a jeho prvá verzia par-skpl-1.0 (https://korpus.sk/skpl.html) bola sprístupnená 3. 12. 2018 v rozsahu takmer 8,2 milióna tokenov (v slovenskej časti 4,12 milióna tokenov, v poľskej časti 4,06 milióna 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.

B. Prvá verzia **slovensko-španielskeho paralelného korpusu** bola sprístupnená 17. 7. 2019 v rozsahu takmer 11,5 milióna 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.

4. Spravovanie a zdokonaľovanie elektronickej verzie základných kodifikačných príručiek schválených MK SR, rozširovanie a aktualizácia ďalších elektronických lingvistických zdrojov.

SPLNENÁ

Realizácia tohto bodu sa v celom projektovom období uskutočňovala v spolupráci s pracovníkmi iných oddelení Jazykovedného ústavu Ľ. Štúra SAV. Postupne boli na slovníkový portál Jazykovedného ústavu Ľ. Štúra SAV (https://slovnik.juls.savba.sk/) pridané elektronické verzie týchto diel: *Ortograficko-gramatický slovník*, tri zväzky *Slovníka slovenských nárečí*, *Historický slovník slovenského jazyka*, *Slovník prepisov z orientálnych jazykov* a *Retrográdny slovník súčasnej slovenčiny*. Slovníkový portál ponúka používateľom aj vygenerované prekladové ekvivalenty zo slovensko-francúzskeho paralelného korpusu. Okrem rozšírenia portálu o nové lexikografické zdroje sa existujúce diela priebežne aktualizujú podľa potreby.

Mesačne bolo v roku 2021 v priemere evidovaných 1 124 455 vyhľadaní konkrétneho slovníkového hesla z 68 839 unikátnych ip adries.

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

SPLNENÁ

A. *MAJCHRÁKOVÁ, Daniela – CHLPÍKOVÁ, Katarína – BOBEKOVÁ, Kristína:* ***Slovník kolokácií prídavných mien v slovenčine.*** *Bratislava: Veda 2017. 344 s. ISBN 978-80-224-1633-7.*

V slovníku sú spracované kolokačné profily 500 najfrekventovanejších prídavných mien vybraných zo Slovenského národného korpusu. Jednotlivé kolokačné profily predstavujú vzorku rôznych typov slovných spojení s prídavným menom, ktoré reálne fungujú v slovenských textoch. Okrem kolokácií – typických spojení a ustálených viacslovných pomenovaní, ktoré tvoria základ kolokačného profilu každého prídavného mena, sú v rámci heslovej state osobitne vyčlenené

frazémy a frázy. Slovník obsahuje aj štatistické prehľady, ktoré umožňujú čitateľovi porovnať spájateľnostný potenciál jednotlivých prídavných mien (najviac slovných spojení obsahuje heslo veľký – 239, najmenej heslo ústavný – 7), porovnať spájateľnosť prídavných mien so štatistickými hodnotami zo Slovenského národného korpusu, prípadne aj so spájateľnosťou podstatných mien. Slovník sa môže využiť ako východisko ďalšieho lexikologického, lexikografického alebo korpusového výskumu, dobrou pomôckou môže byť v prekladateľskej praxi, ako aj pri výučbe slovenského jazyka.

V januári 2019 získalo dielo v medzinárodnej súťaži Slovník roku **Čestné uznanie poroty**.

B. *ŠIMKOVÁ, Mária – GAJDOŠOVÁ, Katarína – KMEŤOVÁ, Beáta – DEBNÁR, Marek:* ***Slovenský národný korpus. Texty, anotácie, vyhľadávania.*** *Bratislava: Jazykovedný ústav Ľ. Štúra SAV – Vydavateľstvo Mikula 2017. 168 s. ISBN 978-80-88814-98-6.*

Kolektívna publikácia mapuje tvorbu, štruktúru a možnosti využitia zdrojov Slovenského národného korpusu. Je určená lingvistom, učiteľom slovenského jazyka a cudzích jazykov na všetkých stupňoch škôl, prekladateľom, študentom a všetkým záujemcom o korpusové a jazykové databázy. V knihe možno nájsť okrem iného prehľad korpusov, podkorpusov a jazykových zdrojov Slovenského národného korpusu, podrobný opis dvoch základných anotácií (morfologickej a štýlovo-žánrovej), terminologický slovník, výber najčastejšie používaných metaznakov a metaoperátorov používaných pri vyhľadávaní, ako aj praktické ukážky (návody) postupov pri vyhľadávaní jazykových javov v rôznych korpusoch.

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

SPLNENÁ

A. *GARABÍK, Radovan – KMEŤOVÁ, Beáta – ŠIMKOVÁ, Mária – ZUMRÍK, Miroslav a kol.:* ***Frekvenčný slovník slovenčiny na báze Slovenského národného korpusu.*** *Bratislava: Veda 2017. 562 s. ISBN 978-80-224-1630-6.*

Frekvenčný slovník slovenčiny na báze Slovenského národného korpusu vychádza po takmer 50 rokoch od prvého vydania slovníka tohto typu na Slovensku (Jozef Mistrík: Frekvencia slov v slovenčine. Bratislava: Vydavateľstvo Slovenskej akadémie vied 1969). Nový frekvenčný slovník zachytáva najfrekventovanejšie slová súčasnej slovenčiny, ako sa používali v rozpätí 25 rokov z prelomu 20. a 21. storočia. Frekvenčné údaje sú založené na textových a jazykových zdrojoch Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV. Frekvenčný slovník slovenčiny na báze Slovenského národného korpusu obsahuje: Slovník lem podľa absolútnej frekvencie (28 456 najfrekventovanejších slov), Slovník lem podľa priemernej redukovanej frekvencie (28 477 najfrekventovanejších slov), Slovník lem podľa abecedy s rozšírenými frekvenčnými údajmi (rovnakých 28 477 najfrekventovanejších slov) a doplňujúce frekvenčné zoznamy: Frekvenčný zoznam interpunkčných znamienok, Frekvenčný zoznam grafických symbolov a číslic, Frekvenčný zoznam grafém, Frekvenčný zoznam grafém stojacich na začiatku slov, Frekvenčný zoznam slovných druhov.

B. *GARABÍK, Radovan a kol.:* ***Retrográdny slovník súčasnej slovenčiny. Slovné tvary na báze Slovenského národného korpusu.*** *Bratislava: VEDA 2018. 848 s. ISBN 978-80-224-1699-3.*

Dielo poskytuje rozsiahle, reprezentatívne a prehľadne spracované údaje o zakončení tvarov slov v slovenčine spolu so štatistickým spracovaním ich výskytov a distribúcie. Ide o prvý slovník svojho druhu, v ktorom sú zahrnuté a systematicky spracované zakončenia slov v slovenčine na základe rozsiahleho korpusového materiálu. Slovník ako praktická referenčná príručka typických i menej častých zakončení slov je určený odbornej i širokej laickej verejnosti. Okrem informácií o používaní zakončení tvarov slov čitateľom poskytuje rýchlu orientáciu vo frekvencii realizácie gramatických kategórií a pomáha odhaľovať neintuitívne javy a vzťahy v jazyku.

V rámci medzinárodnej súťaži Slovník roku 2020 získala táto publikácia **2. miesto v súťaži**

o hlavnú cenu (Slovník roku 2020) a **Cenu poroty**. Zároveň bola ocenená aj **Prémiou Literárneho fondu za vedeckú a odbornú literatúru** za rok 2018.

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

SPLNENÁ

GAJDOŠOVÁ, Katarína – ŠIMKOVÁ, Mária a kol.: ***Frekvenčný slovník hovorenej slovenčiny na báze Slovenského hovoreného korpusu.*** *Bratislava: Veda 2018. 416 s. ISBN 978-80-224-1678-8*

Frekvenčný slovník hovorenej slovenčiny na báze Slovenského hovoreného korpusu predstavuje prvé samostatné spracovanie výskytov slov a spojení v ústnej podobe štandardnej slovenčiny. Kolektív autorov z JÚĽŠ SAV v ňom nadväzuje na Frekvenčný slovník slovenčiny na báze Slovenského národného korpusu (Bratislava, 2017), v ktorom sú spracované výskyty slov v písaných textoch. Obidva slovníky tak poskytujú celistvý obraz o najpoužívanejšej časti slovnej zásoby súčasnej slovenčiny. Vo Frekvenčnom slovníku hovorenej slovenčiny sa nachádzajú výskyty slov a spojení slov v prehovoroch rôzneho druhu z celého územia Slovenska, ktoré sú spracované v Slovenskom hovorenom korpuse v rozsahu takmer 6,6 milióna textových jednotiek. Frekvenčný slovník obsahuje takmer 10 000 najfrekventovanejších slov usporiadaných podľa absolútnej frekvencie, podľa priemernej redukovanej frekvencie a podľa abecedy s rozšírenými frekvenčnými údajmi. V ďalších častiach sa nachádzajú frekvenčné zoznamy dvoj-, troj- a štvorslovných spojení slov, frekvenčné zoznamy slov použitých v osobitných komunikačných situáciách, skratiek a značiek, frekvenčné zoznamy grafém, slovných druhov a štruktúrnych značiek používaných v Slovenskom hovorenom korpuse.

Dielo získalo **hlavnú cenu Literárneho fondu za vedeckú a odbornú literatúru** za rok 2018.

ŠIMKOVÁ, Mária – LEVICKÁ, Jana – DEBNÁR, Marek: ***Dynamické javy v súčasnej slovenčine a jej výskume****. Bratislava: VEDA 2018. 196 s. ISBN 978-80-224-1679-5*

Publikácia predstavuje príspevok ku skúmaniu a opisu zmien v súčasnej slovenčine súvisiacich s vývinom celého jazykového spoločenstva i s novými metódami a možnosťami výskumu jazyka a textových útvarov. Autori sa zamerali na niekoľko oblastí: pravopis alebo adaptácia prevzatých slov, lexikológia vrátane terminológie, štylistika s presahom do dištančného čítania. Premenlivé vlastnosti jazykového systému slovenčiny a sondy do pohybov v parole sú dokumentované na báze dát Slovenského národného korpusu s názornými ukážkami efektívneho využitia korpusových zdrojov a metód korpusovej lingvistiky. Kniha je primárne určená lingvistom, literárnym vedcom, záujemcom zo súvisiacich interdisciplinárnych oblastí (napr. digital humanities), terminológom, prekladateľom a študentom príslušných odborov.

8. Tvorba a skvalitňovanie terminologického portálu a Slovenskej terminologickej databázy vrátane začleňovania termínov podľa potrieb a požiadaviek odboru štátneho jazyka MK SR.

SPLNENÁ

A. **Terminologický portál** bol vytvorený s cieľom poskytnúť prehľad o terminologických zdrojoch a informáciách teoretického aj praktického charakteru. V päťročnom projektovom období bola jeho pilotná verzia kvalitatívne aj kvantitatívne vylepšená. Aktuálne má verejnosť k dispozícii v prvom rade rozsiahly komplex teoretických zdrojov: archív časopisu *Slovenské odborné názvoslovie* (109 súborov vo formáte PDF), archív časopisu *Československý terminologický časopis* (30 súborov vo formáte PDF) a osobitnú databázu terminologických prác profesora Jána Horeckého, ktorá obsahuje celkovo 1 418 bibliografických položiek. Pri väčšine z nich sa nachádza celý text vo formáte PDF, pri 951 je anotácia, kľúčové slová a priradený tematický deskriptor zo Sústavy odborov vedy a techniky a číselníka odborov vedy a techniky podľa smernice č. 27/2006-R vydanej Ministerstvom školstva, vedy, výskumu a športu SR 21. 12. 2006. Terminologické zdroje tvorí predovšetkým Slovenská terminologická databáza, 22 terminologických slovníkov vydaných v rokoch 1952 –

1964 (s elektronickou verziou – 2 511 zdigitalizovaných strán vo formáte PDF) a zoznam približne 400 výkladových a prekladových terminologických slovníkov a encyklopédií zatriedených do príslušných tematických oblastí na základe Sústavy odborov vedy a techniky a číselníka odborov vedy a techniky podľa smernice č. 27/2006-R vydanej Ministerstvom školstva, vedy, výskumu a športu SR 21. 12. 2006. Ku každému lexikografickému či terminografickému dielu bola priradená kratšia anotácia a označenie jazykov spracovaných v príslušnom diele. Vo viacerých prípadoch je anotačný záznam doplnený o popis štruktúry hesla. Na portáli sa nachádzajú aj užitočné linky na zahraničné databázy, terminologické centrá a zdroje, ako aj zoznam terminologických komisií pri ministerstvách a ostatných ústredných orgánoch štátnej správy. Dáta a informácie na terminologickom portáli sa priebežne aktualizujú a dopĺňajú podľa potreby.

B. **Slovenská terminologická databáza** prechádzala v roku 2016 na novú platformu Semantic Media Wiki, preto v roku 2017 nasledovala kontrola, aktualizácia a úprava dát, v rámci ktorej bolo revidovaných 1000 terminologických záznamov z celkového počtu 7 827 verejne sprístupnených terminologických záznamov. Ďalej nasleduje zoznam rozširovania databázy v jednotlivých rokoch:

2017 – takmer **2 500 nových terminologických záznamov** z oblasti astronómie, filozofie, fyziky, informačných technológií a medicíny;
2018 – **748 nových terminologických záznamov** na základe získania autorských práv k 1. dielu Encyklopédie medicíny (Kadlec, O. et al.: Encyklopédia medicíny. I. Bratislava: Asklepios 1993). Celý súbor termínov prešiel odbornou revíziou a aktualizáciou;
2019 – **1069 terminologických záznamov** na základe získania autorských práv k trom dielam z oblasti logistiky a obchodu; ďalších takmer **900 terminologických záznamov** bolo spracovaných z oblasti astronómie, filozofie, fyziky, informačných technológií a medicíny;
2020 – 336 nových terminologických záznamov z oblasti verejnej správy. Okrem tohto plánovaného rozšírenia Slovenskej terminologickej databázy sa vytvoril priestor na doplnenie a revidovanie terminologických záznamov z oblasti medicíny v rozsahu **1 900 nových terminologických záznamov**, na aktualizáciu a sprístupnenie terminologických záznamov z oblasti masmediálnej komunikácie (580 terminologických záznamov) a na niekoľko ďalších menších úprav a doplnení;
2021 – 1005 nových terminologických záznamov v oblasti epidemiológie a imunológie a **550 terminologických záznamov** z oblasti práva. Celkovo bolo aktualizovaných a revidovaných vyše 5 400 terminologických záznamov. Aktualizácie a kontroly sa zároveň realizovali aj v rozpracovanej kategórii IT, kde pribudlo **66 nových terminologických záznamov**.

K februáru 2022 obsahovala Slovenská terminologická databáza vo svojej verejne prístupnej časti **16 324 terminologických záznamov** zaradených do **23 kategórií** – oproti východiskovým 7 827 terminologickým záznamov zaradeným do 17 kategórií ide o nárast viac ako 100 %.

9. Vybudovanie korpusu textov štátnej správy s cieľom extrakcie termínov pre potreby terminologickej analýzy.

SPLNENÁ

Verzia korpusu	Rozsah verzie	Sprístupnenie verzie
gov-web-1.0	11,7 milióna tokenov	3. 12. 2018
gov-web-2.0	12,5 milióna tokenov	22. 7. 2020

Oproti pôvodnému plánu boli sprístupnené až dve verzie tohto špecializovaného korpusu, pričom došlo v druhej verzii k miernemu nárastu dát. Najnovšia verzia korpusu sa skladá z textov štátnych inštitúcií dostupných na webových doménach gov a egov do roku 2020. Rovnako ako ostatné korpusy SNK, aj tento je lematizovaný a morfologicky anotovaný, pri textoch sú uvedené základné informácie o ich url a čase získania. Východiskové texty boli deduplikované na úrovni odsekov.

10. Zhromažďovanie a spracúvanie odborných textov s cieľom extrakcie termínov vybraných vedných odborov pre Slovenskú terminologickú databázu.

SPLNENÁ

Vzhľadom na nižší podiel odborných textov v 6. a 7. verzii hlavného korpusu (11 % a 9,5 %) sa pozornosť v rokoch 2017 – 2021 zamerala na ich získavanie a zhromažďovanie. Náročnosť zhromažďovania tohto typu textov spočíva aj v skutočnosti, že s internacionalizáciou výskumu sa ťažisko vedeckého publikovania presúva do anglického jazyka, čo má negatívne dôsledky nielen na odborný jazyk v slovenčine ako taký, ale aj na slovenskú terminológiu.

Rok	Počet dokumentov	Počet tokenov
2017	1631	vyše 4,7 milióna tokenov
2018	4770	vyše 8,4 milióna tokenov
2019	4026	vyše 12,4 milióna tokenov
2020	2731	vyše 18,4 milióna tokenov
2021	14072	vyše 27,3 milióna tokenov

11. Dobudovanie Slovenského hovoreného korpusu a skvalitňovanie prístupu verejnosti do jeho databáz.

SPLNENÁ

Verzia hovoreného korpusu	Rozsah verzie	Sprístupnenie verzie
korpus s-hovor-6.0	6 593 000 tokenov	30. 11. 2017
korpus s-hovor-7.0	7 800 000 tokenov	marec 2022

Od verzie hovoreného korpusu **s-hovor-6.0** majú značky použité v prepise, ktoré sú používateľom k dispozícii, podobu štruktúrnych značiek a sprístupnila sa aj možnosť vypočuť si príslušnú časť zvukového záznamu priamo vo vyhľadávacom nástroji NoSketch Engine.

Najnovšia verzia hovoreného korpusu **s-hovor-7.0** obsahuje 852 hodín zvukových záznamov. Oproti predchádzajúcej verzii ide o nárast o vyše 100 nahrávok a viac ako 100 hodín zvukových záznamov, resp. o vyše 1,2 milióna tokenov.

12. Dobudovanie Historického korpusu slovenčiny a skvalitňovanie prístupu verejnosti do jeho databázy.

SPLNENÁ

Verzia historického korpusu	Rozsah verzie	Sprístupnenie verzie
korpus hist-5.0	997 809	24. 2. 2020
korpus hist-6.0	zatiaľ neurčený	marec 2022

Verzia korpusu **hist-5.0** v rozsahu takmer 1 milióna. tokenov bola sprístupnená vo februári 2020 a oproti predchádzajúcej verzii obsahuje o vyše 80 tisíc tokenov viac. Bolo do nej zahrnutých 5 pôvodných historických textov (*Poznámky a rady hygienické, prírodopisné, meteorologické, ekonomické*, *Receptár gemerský*, *Receptár nitriansky*, *Receptár osturniansky – zápisy o ľudovom liečení*, *Receptár prešovský*), ktoré sa prepisovali z fotokópií originálov špeciálne pre túto verziu korpusu. Prepisy sa korigovali, zjednocovali a dopĺňali o relevantné informácie v štruktúrnych značkách a komentároch.

Najnovšia verzia tohto korpusu **hist-6.0** sa od predchádzajúcej líši kvantitatívne aj kvalitatívne. V dátach do tejto verzie boli uskutočnené úpravy na troch úrovniach: najskôr išlo o revíziu doterajších textov, aby všetky spĺňali podmienku transliterácie. Ďalšou úpravou bolo zjednotenie a zjednodušenie označenia cudzojazyčných textov a skratiek. Na úrovni vyhľadávania

bola doplnená možnosť výberu rozlišovať alebo nerozlišovať ypsilon od ÿ. Do aktuálnej verzie pribudlo 5 manuálne prepisovaných pamiatok z predspisovného obdobia – 3 administratívno-správne a 2 z oblasti liečiteľstva. Navyše sa do tejto verzie historického korpusu podarilo získať texty z dvoch publikácií APVV projektu č. APVV-16-0374: *Slovaciká z bývalého Uhorského kráľovstva na príklade Horného Uhorska* (1500 – 1780).

13. Rozširovanie a skvalitňovanie Korpusu nárečí a budovanie Archívu nárečí v rámci záchrany kultúrneho dedičstva – digitalizácie autentických prehovorov autochtónnych nositeľov nárečí, ako aj v súlade s potrebami tvorby Slovníka slovenských nárečí a slavistických výskumov.

SPLNENÁ

A. rozširovanie Korpusu nárečí

Verzia nárečového korpusu	Rozsah verzie	Sprístupnenie verzie
korpus dialekt-4.0	711 766 tokenov	18. 12. 2018
korpus dialekt-5.0	zatiaľ neurčený	marec 2022

Štvrtá verzia Korpusu nárečí Slovenského národného korpusu, ktorý sa v rámci SNK postupne buduje od roku 2013, bola oproti predchádzajúcej dialekt-3.0 obohatená o 28 nových korpusovo spracovaných textových zdrojov, čo predstavovalo nárast o vyše 43 %. V rámci skvalitňovania korpusu došlo k zjednoteniu názvov zdrojov (doc.source). Zoznam spracovaných zdrojov v korpuse **dialekt-4.0** je dostupný aj s plnými bibliografickými údajmi. Korpus nárečí nie je lematizovaný ani morfologicky anotovaný, vyhľadáva sa v ňom na základe konkrétneho slova (word) a pomocou zástupných (meta)znakov. Pri prepisoch sú uvedené sociolingvistické údaje o informátoroch a explorátoroch, ako aj informácie o pôvode a obsahu nahrávky.

Do najnovšej, piatej verzie tohto korpusu sa pripravovali dáta v roku 2021 vychádzajúce z naskenovaných a rekonštruovaných zdrojov (celkovo 51 zdrojových textov) rozličného rozsahu, s dátami z predchádzajúcej verzie korpusu tak nová verzia obsahuje cez 100 zdrojových textov. Zväčša ide o ukážky nárečových textov z popularizačných textov alebo vlastivedných monografií, ale aj nárečové texty viacerých dialektológov, napr. J. Dudášovej, G. Horáka, E. Jónu, V. Kováčovej.

B. Budovanie Archívu nárečí

V oddelení SNK sa od roku 2013 systematicky zbierajú a uchovávajú pôvodné zvukové záznamy nárečových prehovorov z rôznych období a lokalít, ktoré predstavujú jedinečné kultúrne dedičstvo. Získané nahrávky, ktoré boli roztrúsené na jednotlivých slovakistických pracoviskách slovenských univerzít alebo u individuálnych bádateľov, sa na pracovisku SNK digitalizovali, technicky spracúvali do kvalitnejšej podoby (odšumovali a pod.) a pripisovali sa k nim existujúce metadáta (informácie o hovoriacich a nahrávajúcich). K niektorým zvukovým záznamom existujú čiastočné alebo úplné prepisy, časti z nich sa v niekoľkých prípadoch stali súčasťou postupovej práce alebo boli knižne vydané. Uvedené informácie sa v oddelení SNK zapisovali do databázy, ktorá bola prvýkrát zverejnená v roku 2017 a momentálne je k dispozícii jej tretia verzia.

Verzia databázy	Veľkosť databázy	Sprístupnenie databázy
1. verzia databázy	100 nahrávok	28. 8. 2017
2. verzia databázy	280 nahrávok	18. 12. 2017
3. verzia databázy	333 nahrávok	20. 12. 2018

V sprístupnenej časti interaktívnej databázy (http://korpus.sk/dialect_recordings.html) sa nachádzajú údaje o 333 nárečových nahrávkach v celkovej dĺžke trvania viac ako 74 hodín od 11 individuálnych a 7 inštitucionálnych poskytovateľov. Záujemcovia môžu pracovať

s vybranými alebo všetkými zdrojmi prezenčne v priestoroch SNK JÚĽŠ SAV po dohode s pracovníkmi SNK. Zdroje Archívu nárečí SNK sú určené predovšetkým na dialektologický výskum, ale môžu poslúžiť aj historikom, etnológom, kulturológom a širokej odbornej verejnosti.

14. Koncepcia a realizácia sémantickej anotácie korpusu (identifikácia viacslovných pomenovaní, ručná anotácia pomenovaných jednotiek, budovanie ontológií).

SPLNENÁ

V rámci riešenia tejto úlohy v SNK v rozmedzí rokov 2018 – 2021 bola sémantická analýza textov vymedzená ako anotácia pomenovaných entít (meno osoby, názov obce, inštitúcie bez ohľadu na počet slov). Pri anotácii sa uvedeným entitám ručne 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. Proces ručnej anotácie je uľahčený automatickou predanotáciou, založenou na využití kolekcie lexikónov pomenovaných entít pre identifikované kategórie, 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 anotáciu boli využité texty z korpusu slovenskej Wikipédie, pričom pri ich výbere sa prihliadalo na primeraný rozsah a rovnomerné zastúpenie tém z hľadiska obsahu. Do súčasného stavu bolo zanotovaných 299 textov, v ktorých bolo označených viac ako 232 tisíc pomenovaných entít. Anotácia každého textu bola realizovaná dvoma anotátormi. Následná korekcia nezrovnalostí v nimi označených pomenúvajúcich entitách analyzovala a vyriešila tretia osoba.

Na podmnožine súborov textov s anotáciami bol natrénovaný efektívny model pre automatické rozpoznávanie pomenúvajúcich entít s použitím nástroja NameTag 1. Tento model spolu s dátovými množinami a údajmi o úspešnosti bude sprístupnený pod slobodnou verejnou licenciou.

15. Sprístupňovanie jednotlivých korpusov a elektronických databáz na internete na vedecko-výskumné a učebné využitie pre slovenských i zahraničných bádateľov, zabezpečovanie korpusových databáz nástrojmi vhodnými na počítačové spracovanie prirodzeného jazyka, na výučbové, lingvistické a iné vedeckovýskumné využitie, príprava a realizácia školení pre záujemcov o používanie korpusových zdrojov, organizovanie medzinárodných konferencií Slovko.

SPLNENÁ

Sprístupňovanie korpusových databáz a ich skvalitňovanie vhodnými nástrojmi je popísané v predchádzajúcich bodoch, preto sa v tejto časti obmedzíme na súhrn informácií týkajúcich sa organizovania bienálnej konferencie a realizácie školení o používaní korpusových zdrojov.

A. *Zorganizovanie troch ročníkov medzinárodnej konferencie SLOVKO*

9. ročník SLOVKO 2017 sa uskutočnil 25. – 27. októbra 2017 v Bratislave na tému počítačového spracovania prirodzeného jazyka, korpusovej lingvistiky, terminológie, e-terminológie a e-terminografie, e-lexikografie a gramatického korpusového výskumu. Celkovo sa ho zúčastnilo 62 účastníkov zo SR, Česka, Gruzínska, Chorvátska, Fínska, Nemecka, Poľska, Rakúska, Ruska, Švédska a Ukrajiny. Zborník z tohto podujatia vyšiel v špeciálnom čísle Jazykovedné časopisu č. 2 a obsahoval 31 príspevkov v anglickom jazyku.

10. ročník SLOVKO 2019 sa uskutočnil 23. – 25. októbra 2019 v Bratislave na tému počítačového spracovania prirodzeného jazyka, korpusovej lingvistiky, dynamiky jazyka a jazykových zmien. Celkovo sa ho zúčastnilo 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. Zborník z tohto podujatia vyšiel v špeciálnom čísle Jazykovedné časopisu č. 2 a obsahoval 33 príspevkov.

11. ročník SLOVKO 2021 sa uskutočnil 13. – 15. októbra 2021 v Bratislave na tému počítačového spracovania prirodzeného jazyka, korpusovej lingvistiky a interdisciplinárneho výskumu. Celkovo sa ho zúčastnilo vyše 60 účastníkov, z toho viac ako 20 online, z rôznych pracovísk SR a z Česka, Ruska, Slovinska, Rakúska, Nemecka, Švédska, Ukrajiny, Chorvátska. Zborník z tohto podujatia vyšiel v špeciálnom čísle Jazykovedné časopisu č. 2 a obsahoval 34 príspevkov v anglickom jazyku.

B. *Príprava a realizácia školení pre záujemcov o používanie korpusových zdrojov.*

2017	8 praktických seminárov (celkovo 132 účastníkov)
2018	15 praktických seminárov (celkovo 227 účastníkov)
2019	12 praktických seminárov vrátane 4 prezentácií (takmer 350 účastníkov)
2020	3 praktické semináre (1 organizovaný online)
2021	7 praktických seminárov (celkovo 32 účastníkov)

C. *Počet registrovaných používateľov korpusových databáz*

2017	597
2018	590
2019	727
2020	669
2021	673

16. Analýza korpusového materiálu so zameraním na prídavné mená a ich opis.

ČIASTOČNE SPLNENÁ

A. *Koncepcia a príprava publikácie Tvary prídavných mien s korpusovými príkladmi.*

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, 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 (publikácia vydaná v roku 2016 v predchádzajúcej etape), 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. V novom projektovom období sa plánuje pokračovať v tejto úlohe a zvŕšiť ju samostatnou publikáciou.

B. *Príprava a vydanie monografie Lexikálno-gramatický opis adjektív v písanej a hovorenej slovenčine na báze korpusového materiálu*

Pri príprave tejto úlohy boli identifikované také typy a javy súvisiace s prídavnými menami, ktoré neboli doteraz v centre pozornosti, napr. deminutívne a augmentatívne tvorenie prídavných mien, spájateľnosť a pragmatické aspekty fungovania prídavných mien z hľadiska zmien v spoločnosti a dynamiky jazyka či analýza synonymných vzťahov vo vybranej skupine adjektív v odborných textoch. Tieto témy boli predstavené na konferenciách doma i v zahraničí. V rokoch 2018 – 2021 bolo vypracovaných a publikovaných celkovo 8 štúdií, z toho 2 v anglickom jazyku.

V roku 2018:

Levická, J.: Konkurenčnosť synonymných adjektív v lekárskej terminológii.

V roku 2019:

Gajdošová, K.: Vzťahy medzi vybranými augmentatívnymi adjektívami.

Gajdošová, K.: Sufixálne augmentatívne adjektíva v súčasnej slovenčine.

Krolčíková, R.: Fungovanie adjektív vo vybraných dielach Dušana Mitanu (Analýza na báze Slovenského

národného korpusu).
Šimková, M.: Deminutívne adjektíva v súčasnej slovenčine (na báze Slovenského národného korpusu).
Šimková, M.: Minulé a súčasné kontexty prídavného mena *politický*.

V roku 2021:

Levická, J.: Usage and Empirical Productivity of International Adjectival Suffixes in Slovak Based on General and Specialised Corpora.
Levická, J.: Usage and empirical productivity of international adjectival suffixes in Slovak revisited.

Opatrenia súvisiace s pandémiou COVID-19 a personálne zmeny v oddelení neumožnili dokončenie tejto úlohy v zamýšľanom rozsahu a podobe.

17. Vybrané priebežné úlohy podľa Harmonogramu zmluvy

A. *Dopĺňanie a) všeobecného korpusu, b) paralelných korpusov a c) ďalších korpusov SNK aktuálnymi textami*

Rok	Počet digitalizovaných strán	Počet zrekonštruovaných strán
2017	25 000	25 000
2018	30 500	22 000
2019	20 000	22 000
2020	9 790	12 497
2020	12 000	10 450

B. *Správa a aktualizácia prezentačnej stránky SNK na Facebooku*

Snaha o propagáciu výstupov oddelenia Slovenského národného korpusu sa nielen z dôvodu protiepidemických opatrení rozširovala predovšetkým v online priestore, čo sa odráža aj na zvýšenom počte príspevkov a najmä sledovateľov v porovnaní s rokom 2017.

2017 1009 sledovateľov
2018 1110 sledovateľov
2019 1255 sledovateľov
2020 1522 sledovateľov
2021 1666 sledovateľov

Úlohy nad rámec špecifikácie predmetu Zmluvy

1. Tvorba a rozširovanie korpusu slovenských textov z Wikipédie a Necyklopédie

Verzia korpusu	Rozsah korpusu	Sprístupnenie
Korpus wiki-2017-02	45,1 mil. tokenov	2017
Korpus wiki-2018-03	47,2 mil. tokenov	2018
Korpus wiki-2019-08	50,1 mil. tokenov	2019

Posledná, šiesta verzia tohto korpusu obsahuje slovenské texty z Wikipédie dostupné k 1. 8. 2019 a neobsahuje niekoľko chýb spracovania značiek MediaWiki z predchádzajúcich verzií. Korpus je lematizovaný (s rozlíšením malých a veľkých začiatočných písmen pri všeobecných a vlastných pomenovaniach) a morfologicky anotovaný. Pri textoch je uvedená informácia o ich zdroji.

2. Vytvorenie slovensko-rumunského paralelného korpusu

Prvá verzia **par-skro-fic-1.1** bola sprístupnená 24. augusta 2017 ako malý experimentálny korpus v rozsahu takmer 1,3 mil. tokenov (603 111 tokenov v slovenskej časti, 688 867 tokenov v rumunskej časti). Slovensko-rumunský paralelný korpus obsahuje preklady troch literárnych textov

z rumunčiny do slovenčiny a jedného dokumentu o vzájomnej spolupráci. Texty sú automaticky zarovnané po vetách. Slovenské texty sú automaticky morfologicky anotované tagerom MorphoDiTa natrénovaným v SNK na báze tagsetu vypracovaného v Slovenskom národnom korpuse, rumunské texty sú anotované TreeTaggerom.

3. Vytvorenie špecializovaného korpusu textov súdnych rozhodnutí

Špecializovaný korpus textov súdnych rozhodnutí **od-justice-1.0**, sprístupnený 7. 12. 2018 v rozsahu vyše 4 miliardy tokenov, bol vytvorený z textov dostupných v rámci projektu OpenData sprístupnených Ministerstvom spravodlivosti Slovenskej republiky.

Korpus je lematizovaný a morfologicky anotovaný, pri textoch je uvedená základná informácia o url, o čase a mieste ich vzniku.

4. Vytvorenie špecializovaného referenčného korpusu

Referenčný korpus **prim-7.0-frk** 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.* Takisto poslúžil aj ako zdrojové východisko pre *Retrográdny slovník súčasnej slovenčiny – slovné tvary na báze Slovenského národného korpusu.*

Rozsah korpusu je 253 137 609 tokenov, celkový objem doň zahrnutých textov predstavuje 158 281 dokumentov. Korpus je lematizovaný a morfologicky anotovaný na základe tagsetu SNK, na anotáciu bol použitý tager MorphoDiTa s osobitným natrénovaním na rozpoznávanie vlastných mien.

Plnenie úlohy podľa Dodatku č. 1 Zmluvy z roku 2018

Zmluvné strany sa v tomto dodatku dohodli na vydaní 2 publikácií:

ŠIMKOVÁ, Mária – GAJDOŠOVÁ, Katarína: ***Slovenský národný korpus. Používanie, príklady, postupy.*** *Bratislava: Jazykovedný ústav Ľ. Štúra SAV – Vydavateľstvo Mikula 2020. 336 s. ISBN 978-80-99987-00-6.*
Cvičebnica nadväzuje na publikáciu Slovenský národný korpus Texty, anotácie, vyhľadávania, je určená lingvistom, učiteľom slovenského jazyka a cudzích jazykov na všetkých stupňoch škôl, prekladateľom, študentom a všetkým záujemcom o korpusové a jazykové databázy. Rámcovým cieľom autoriek je predstavenie čo najviac spôsobov vyhľadávania v istej postupnosti aj podľa poznaných potrieb a daností doterajších používateľov a zároveň poukázať na pestrosť jazyka a jeho dynamiku. Publikácia umožňuje predstaviť možnosti využitia korpusov v školskej praxi pri príprave konkrétnych úloh alebo ukážok preberaného učiva.

Človek a jeho jazyk 4. ***Terminologické inšpirácie profesora Jána Horeckého/Man and His Language 4. Selected Terminological Papers of J. Horecký.*** *J. Levická – M. Zumrík (red./eds.). Veda, vydavateľstvo Slovenskej akadémie vied Bratislava 2019, 400 s., ISBN 978-80-224-1740-2.*
Publikácia zo série *Človek a jeho jazyk* sa usiluje predstaviť dielo Jána Horeckého ako zdroj inšpirácií pri ďalšom skúmaní jazyka a terminológie v historickom i systémovom aspekte. Do tohto výberu zostavovatelia zaradili terminologické práce J. Horeckého z rokov 1955 – 1997, v ktorých sa zamýšľal nad kľúčovými terminologickými a onomaziologickými otázkami, naznačil možnosti interlingválnej terminologickej komparácie a bilancoval ustaľovanie slovenskej terminológie. Publikáciu uvádzajú štúdie a texty oboch zostavovateľov. Špecifickosťou tohto zväzku je bilingválne slovensko-anglické spracovanie, ktorého cieľom je sprostredkovanie terminologického uvažovania a prác profesora Jána Horeckého nielen slovenskému, ale aj zahraničnému čitateľovi.

Vypracovala: Jana Levická
vedúca oddelenia Slovenského národného korpusu