SLOVENSKÝ NÁRODNÝ KORPUS A KORPUSOVÁ LINGVISTIKA NA SLOVENSKU PO ROKU 2002

Mária Šimková
Jazykovedný ústav Ľudovíta Štúra SAV, Panská 26, 813 64 Bratislava
e-mail: marias@korpus.juls.savba.sk

ŠIMKOVÁ, M.: Slovak National Corpus and Corpus Linguistics in Slovakia after 2002. In: Slovak Speech, 2013, vol. 78, no. 6, p. 354 – 367.

Abstract: In the early 1990s the field of corpus linguistics was not involved within the Slovak linguistics. J. Mistrík (Encyklopédia jazykovedy, 1993) predicted a substantial influence of various factors on the Slovak language and its research. Factors such as computerization, development of corpora and language technologies have also influenced the formation of the Slovak National Corpus. Internal conditions – necessity of material for language research and compiling a new monolingual dictionary of Slovak have also great impact on its establishment. The Slovak National Corpus was founded in 2002 with the support of the Ministry of Education, the Ministry of Culture and the Slovak Academy of Sciences. SNC comprises of several projects primarily focused on linguistic research and language teaching. Currently, the corpus prim-6.1 contains 829 million tokens. Documents – texts in the corpus keep rich metadata description, including detailed style and genre annotation, and morphological annotation. There are many related corpora, e.g. manually morphologically annotated corpus, Slovak WebCorpus, Corpus of legal texts, Corpus of Spoken Slovak, several parallel corpora (Slovak-French, Slovak-Russian, Slovak-Czech, Slovak-English, Slovak-Latin). Separate projects are the Slovak morphology analyzer, Slovak Terminology Database, Slovak WordNet, Corpus of Historical Slovak. The Slovak language resources are quite sufficient for basic language research, but NLP for Slovak requires further support.
Key words: corpus, corpus linguistics, language resources, natural language processing, specialized corpora

V r. 1993 vyšlo vo Vydavateľstve Obzor v Bratislave desaťročie pripravované kolektívne dielo Encyklopédia jazykovedy. Na tvorbe tohto rozsiahleho textu (na 513 stranách je spracovaných vyše 2000 hesiel) sa podieľalo 68 autorov a lektorov z rôznych oblastí jazykovedy a súvisiacich interdisciplinárnych odborov. Heslá ***Slovenský národný korpus, korpusová lingvistika, počítačové spracovanie prirodzeného jazyka*** sa v ňom však nenachádzajú, keďže prvé dve entity v tom čase na Slovensku neexistovali a tretia sa chápala v podstate ako doména informatiky. Z príbuzných disciplín a pojmov sú spracované heslá *matematická lingvistika* (na Slovensku sa jej venoval najmä J. Horecký), *komputačná – strojová lingvistika*, o ktorej sa v Encyklopédii o. i. uvádza, že sa stáva „zložkou kognitívnej vedy a hľadá možnosti uplatňovať poznatky jazykovedy vo výskume umelej inteligencie“, a heslo ***korpus*** s opisom: „ohraničený súbor jazykových výpovedí zaznamenaných písomne, príp. na magnetofónovej páske al. na platni, tvoriacich materiál na výskumnú prácu; množina textov slúžiacich ako základ lingvistického opisu a argumentácie. Pretože je ohraničený, nemôže obsahovať všetky jazykové javy.“

Zostavovateľ Encyklopédie jazykovedy J. Mistrík v úvodnej stati v časti o perspektívach slovenčiny konštatuje: „Slovenčina je prirodzený a živý jazyk … Nesie stopy rozličných vývinových etáp spoločnosti, preto možno predpokladať, že aj na tomto jazyku, ako na každom inom, sa odrazia stopy vývinu budúcich generácií. … V súčasnej situácii nemožno presne predpovedať, ako sa bude slovenčina vyvíjať a rozvíjať, keďže nevieme, ako sa bude vo svete vyvíjať veda a technika a aký osud bude mať tento vývin v našich podmienkach. Máme predstavy o tom, aké sú možnosti rozvoja nášho jazyka, ale nemôžeme mať celkom jasný obraz o faktoroch, ktoré budú tento proces ovplyvňovať. … Náš jazyk stojí dnes na úrovni najmodernejších jazykov sveta … Ďalší ... vývin bude determinovaný situáciou zvonku a možnosťami zvnútra“ (Encyklopédia jazykovedy, 1993, s. 46).

S odstupom dvoch desaťročí od vydania Encyklopédie môžeme o naznačených súvislostiach povedať, že veda a technika sa rozvíjali a ďalej napredujú obrovským tempom, že táto skutočnosť umožňuje nebývalú výmenu informácií, čo ovplyvňuje jazykovú dynamiku, že v rámci globalizácie a internacionalizácie sa národy a jazyky dostávajú do nezvyčajne rozmanitých a intenzívnych kontaktov, a to všetko nevyhnutne ovplyvňuje aj vývin slovenského jazyka a jeho výskum. Na objektívnu analýzu toho, aký osud má celý tento vývin v našich podmienkach, nemáme azda zatiaľ dostatočný odstup. No o predpovedanej determinácii „situáciou zvonku a možnosťami zvnútra“ vieme, že sa osobitne prejavila napríklad pri zakladaní a prejavuje sa aj v súčasnej činnosti a smerovaní pracoviska Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV v Bratislave (http://korpus.sk) a pri elektronizácii jazykovedného výskumu a rozvoji korpusovej lingvistiky na Slovensku.

Vonkajšie faktory – všeobecný rozmach technologizácie a informatizácie, rozvoj korpusov a korpusovej lingvistiky vo svete, vývoj jazykových technológií a strojového prekladu a naliehavá potreba ich zabezpečenia pre slovenčinu v procese prístupových konaní Slovenskej republiky pred prijatím do Európskej únie spôsobovali a postupne zvyšovali tlak na založenie pracoviska počítačovej a korpusovej lingvistiky na Slovensku. Vznik SNK JÚĽŠ SAV de iure sa datuje uznesením vlády SR č. 137 z 13. februára 2002, ktorým sa na základe iniciatívy Ministerstva kultúry SR a s finančnou podporou Ministerstva školstva SR a Predsedníctva SAV schválil ***Projekt vybudovania Národného korpusu slovenského jazyka a elektronizácie jazykovedného výskumu v rokoch 2002 – 2006***. Fungovanie oddelenia a riešenie projektu sa začalo de facto 1. apríla 2002, keď bola jeho vybudovaním a vedením poverená M. Šimková. To, že takéto pracovisko sa stalo novým oddelením Jazykovedného ústavu SAV a nie napr. Ústavu informatiky SAV alebo niektorej z prírodovedne či technicky orientovaných fakúlt Univerzity Komenského, ovplyvnili vnútorné faktory, medzi ktoré patrí predovšetkým lexikografická tradícia v JÚĽŠ SAV obsahujúca okrem významného teoretického zázemia aj skúsenosť s budovaním a využívaním

materiálových zdrojov na opis jednotlivých foriem slovenského jazyka vrátane pilotného interného korpusu textov (bližšie napr. Šimková, 2008) a príprava nového výkladového Slovníka súčasného slovenského jazyka. V Koncepcii starostlivosti o štátny jazyk Slovenskej republiky schválenej vládou SR v r. 2001 bolo preto za jednu z hlavných úloh v oblasti jazykovedy a jazykového výskumu stanovené vybudovanie Národného korpusu slovenského jazyka. Charakteristika korpusu uvedená v Encyklopédii jazykovedy sa postupne stávala neaktuálnou.

Aké bolo postavenie, cieľ a aké sú výsledky a prínosy najmladšieho oddelenia JÚĽŠ SAV v kontexte slovenskej lingvistiky a svetovej korpusovej lingvistiky uplynulých dvoch desaťročí?

Prudký vzostup elektronických jazykových databáz a korpusovej lingvistiky zhruba od 60. rokov v USA (H. Kučera, F. N. Francis – prvý elektronický korpus), V. Británii (R. Quirk a i. – Survey of English Usage) a v ďalších krajinách sa na Slovensku (najmä v JÚĽŠ SAV a v Informačnom centre SAV, neskôr v Laboratóriu počítačovej lingvistiky PdF UK) začal zachytávať na prelome 90. rokov 20. st. Teoretické koncepčné východiská (formulované predovšetkým J. Horeckým) sa začali sľubne napĺňať v praktických pokusoch o generovanie slovenčiny (Páleš, 1994) i v zbieraní a spracúvaní textov v internej elektronickej databáze JÚĽŠ SAV (podrobnejšie Šimková, 2006, 2008; Šimková – Garabík, 2013). Interný korpus, hoci neveľkého rozsahu a bez lingvistickej anotácie, sa od jeho sprístupnenia pracovníkom JÚĽŠ SAV hneď využíval v lexikografickej práci – v nových vydaniach Krátkeho slovníka slovenského jazyka, Pravidiel slovenského pravopisu, pri tvorbe koncepcie a koncipovaní prvých hesiel Slovníka súčasného slovenského jazyka. Rozvíjali sa aj potrebné kontakty a kontrakty so zahraničím (Jarošová, 2001). Celý dobre rozbehnutý proces tvorby a využívania internej textovej databázy však stále viac brzdil nedostatok financií a pracovných kapacít. Voľne dostupné programy na spracovanie textov už kapacitne nezvládali narastajúce množstvo dát a na kvalitnejší softvér boli potrebné finančné prostriedky, ktoré však JÚĽŠ SAV nemal k dispozícii a ani sa nedali získať z bežných zdrojov, napr. z grantovej agentúry VEGA. Na prelome tisícročí, keď sa v okolitých krajinách sprístupňovalo na internete vyhľadávanie v jednotlivých národných korpusoch, sa podobná snaha na Slovensku dostávala do útlmu, zaostávanie za svetom v tejto oblasti narástlo prakticky na 40 rokov a prehlbovalo sa vo všetkých súvisiacich zložkách: na slovenských vysokých školách sa nevyučovala počítačová a korpusová lingvistika, nerobili sa potrebné formálne opisy jazyka, neexistovali vhodné počítačové nástroje pre slovenčinu a pod. Vznikajúce nové pracovisko SNK nemalo k dispozícii žiadne textové dáta – existujúca interná databáza sa nemohla prevziať, pretože korpusové využitie získaných textov prostredníctvom siete internet nebolo právne ošetrené, no mohlo nadviazať na všetky dovtedy získané skúsenosti doma i vo svete.

Z domácich zdrojov sa čerpali najmä informácie o potrebách jednotlivých oblastí slovenského lingvistického výskumu z hľadiska pokrytia jazykových variet, rozsahu, rozmanitosti a kvality spracúvaných textov a ich vonkajšej a vnútornej, lingvistickej anotácie. Dôležitou bola spätná väzba od pracovníkov oddelenia súčasnej lexikológie a lexikografie JÚĽŠ SAV (porov. napr. Šimková, 2013), neskôr i od používateľov z Prešovskej univerzity v Prešove (najmä M. Sokolová a jej žiaci). Pri príprave pravidiel anotácie sme sa mohli oprieť o už takmer uzavretú diskusiu o tejto otázke a o Leechove zásady anotácie: a) možnosť vylúčenia, skrytia anotácie, b) oddeliteľnosť anotácie od textu, c) zrozumiteľnosť anotácie každému používateľovi, d) jasný spôsob a autorstvo anotácie, e) anotácia ako užitočná pomôcka pri analýze, nie bezchybný opis, f) konsenzuálnosť pravidiel anotácie vo vzťahu k rôznym jazykovým teóriám, g) nijaký anotačný systém nie je a priori štandardom, postupne môžu vzniknúť viaceré štandardy (Leech, 1993). Z okolitých, najmä 4 českých pracovísk podobného zamerania prichádzala pomoc v podobe pracovných pobytov, nezištného poskytnutia interných manuálov i nástrojov na spracovanie textov a vyhľadávanie v korpuse, aj v podobe poučení zo zistenia nesprávnosti či nevhodnosti niektorých postupov a pravidelného prednáškového hosťovania na pôde SNK (časť prednášok je zhrnutá v publikácii Insight into the Slovak and Czech Corpus Linguistics, 2006).

Postavenie oddelenia SNK však nebolo jednoduché vzhľadom na niektoré okolnosti, napríklad:

- išlo o najmladšie oddelenie JÚĽŠ SAV nielen v zmysle času založenia, ale aj zloženia pracovného kolektívu, ktorého vekový priemer v tom čase sotva dosahoval 30 rokov (tento priemer prekračovali iba 2 pracovníci) a ktorého členovia boli v oblasti korpusovej lingvistiky prevažne úplnými začiatočníkmi,
- do čisto lingvistického prostredia vstúpili počítačoví odborníci, počítačové metódy a myslenie,
- lingvistická obec stála pred výzvou zvládnuť nové metódy práce s materiálom a po období, keď jazykových dokladov nebolo nikdy dosť, sa bolo treba naučiť selektovať a efektívnejšie analyzovať stále väčšie množstvá dát a formulovať vyhľadávacie príkazy tak, aby sa z materiálu získali všetky požadované informácie,
- autorom a majiteľom autorských alebo distribučných práv bolo potrebné vysvetľovať, že texty poskytnuté do národného korpusu nebudú komerčne ani inak zneužité, komunikácia od oslovenia po získanie súhlasu bola neraz veľmi zdĺhavá a náročná,
- oddelenie sa nachádzalo v osobitnom režime financovania, ktoré bolo (a stále je) prísne podmienené plnením stanovených úloh,
- na obrovské pracovné nasadenie vplývali aj potreby lingvistickej obce a ďalších záujemcov o prácu so slovenskými jazykovými dátami, očakáva-

nia čo najrýchlejšieho priblíženia sa úrovni ostatných národných korpusov a sledovanie rámcového princípu pri tvorbe korpusu: *more data are better data* (čím viac dát, tým sú to lepšie dáta).

Pri schvaľovaní Projektu vybudovania Národného korpusu slovenského jazyka a elektronizácie jazykovedného výskumu v rokoch 2002 – 2006 bolo ako všeobecný cieľ stanovené zachytenie jazyka v celej jeho šírke (novinové texty, beletria, odborné publikácie, hovorený jazyk a pod.) na základe lingvisticky zdôvodnených kritérií a vytvorenie objektívneho a autentického zdroja jazykovej informácie, ktorý by bol materiálovým východiskom na všestranný jazykovedný výskum, tvorbu základných akademických diel a aktualizáciu jestvujúcich praktických jazykových príručiek (porov. materiál predložený na rokovanie vlády SR, dostupný na: http://www.rokovania.sk). V konkrétnych bodoch, ktoré sa dopĺňali o ďalšie úlohy v rámci projektu ***Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu*** Štátneho programu výskumu a vývoja ***Aktuálne otázky rozvoja spoločnosti*** (2003 – 2006, hlavná riešiteľka M. Šimková), to znamenalo:

- koncepčne pripraviť štruktúru korpusu,
- vybudovať vlastný národný korpus,
- vyvinúť softvérové nástroje na spracovanie a správu korpusových databáz,
- koncepčne pripraviť a realizovať segmentáciu, anotáciu, indexáciu získaných textov,
- pripraviť právne podloženú zmluvu na získanie a použitie textov, oslovovať potenciálnych poskytovateľov,
- koncepčne pripraviť a vybudovať:
  - terminologickú databázu,
  - databázu lexikografických diel,
  - paralelný korpus,
  - diachrónny korpus,
  - korpus nárečových textov,
  - korpus hovorených prejavov.

Uvedené koncepčné práce a výstupy bolo možné realizovať len po mnohých jazykových a počítačových analýzach a testovaniach. Na počítačové spracovanie slovenčiny bolo potrebné vyvinúť vlastné metódy a nástroje, každá prehliadnutá osobitosť jazyka sa vo výstupe prejavila ako chyba, ktorú bolo treba nájsť a vyriešiť v celom systéme vzťahov. Jednotlivé zdroje sa pripravovali na nekomerčné, vedecko-výskumné a učebné využitie tak, aby každý záujemca mohol získavať z korpusových dát potrebné jazykové informácie. Vzhľadom na autorský zákon a licenčné zmluvy na poskytnutie textov a ich využitie v korpusovom spracovaní sa pôvodné texty ani korpus ako celok neposkytujú tretím osobám.

Na konci prvej etapy projektu bolo zrejmé, že malá skupina pracovníkov SNK (7 pracovných miest) síce vytvorila veľké dielo, splnila, ba prekročila viaceré základné ciele, no nemohla splniť v potrebnom rozsahu všetky úlohy, najmä tvorbu hovoreného, nárečového a historického korpusu, pri ktorých bola nevyhnutná spolupráca s odborníkmi z príslušných oblastí. Výsledky dosiahnuté v prvej etape boli cenné aj ocenené (***Cena SAV za budovanie infraštruktúry pre vedu***, 2005), pracovisko SNK preklenulo priepasť zaostávania za akceleráciou korpusov a korpusovej lingvistiky vo svete, no v tejto fáze sa ocitlo len kúsok za jej okrajom a na dosiahnutie úrovne dynamicky napredujúcich počítačových a korpusových pracovísk v mnohých krajinách bolo potrebné pokračovať v začatom diele. Prvá etapa sa tak stala dobrým východiskom na prípravu ďalších materiálových zdrojov a počítačových nástrojov pre potreby modernej slovenskej lingvistiky a počítačového spracovania slovenčiny, ktorej pokračovanie bolo možné vďaka podpore MK SR, MŠ SR a SAV aj v ďalších etapách.

Počet pracovných miest v oddelení SNK sa podarilo v každej etape rozšíriť o jedno, takže od r. 2012 má kolektív 9 (prepočítaných) úväzkov (napr. Ústav Českého národného korpusu disponuje približne 40 zamestnancami). Vzhľadom na rozsah stanovených úloh bolo nevyhnutné zapojiť do rôznych počítačových, lingvistických a iných čiastkových prác aj externých spolupracovníkov. Od r. 2004 ich bolo celkovo vyše 250 – niektorí sa podieľali na veľmi špecifických prácach jednorazovo, väčšina však pravidelne niekoľko mesiacov alebo rokov.

Z výsledkov práce kolektívu SNK a spolupracovníkov sa na prvom mieste zvyčajne uvádzajú všeobecný jednojazyčný korpus písaných textov slovenského jazyka – vlastný, primárny Slovenský národný korpus (***prim***) a korpus zvukových záznamov prehovorov v štandardnej slovenčine – Slovenský hovorený korpus (***s-hovor***). Časovú os sprístupňovania jednotlivých verzií týchto korpusov širokej verejnosti na internete a zväčšovanie ich rozsahov znázorňujú súhrny v nasledujúcich tabuľkách. Rovnaký prehľad verzií s-hovor sa nachádza v štúdii Gajdošová – Šimková, 2013, kde sú aj podrobnejšie výklady o zložení hovoreného korpusu, spôsobe prepisu zvukových záznamov, ich anotácii a možnostiach vyhľadávania jazykových a rečových javov v hovorenom korpuse.

Tabuľka č. 1: Verzie Slovenského národného korpusu a ich rozsahy (token = znak alebo reťazec znakov medzi dvoma medzerami, teda slovo, interpunkčné znamienko, matematický znak, grafická značka a pod.)

Názov korpusu	**Prístupný od**	**Počet tokenov (v mil.)**	**Počet slov (v mil.)**	**Počet unikátnych slov (v mil.)**	**Počet unikátnych tvarov slov (v mil.)**
prim0.1	8/2003	30			
prim0.2	12/2003	170			
prim1	2004	182			

Názov korpusu	Prístupný od	Počet tokenov (v mil.)	Počet slov (v mil.)	Počet unikátnych slov (v mil.)	Počet unikátnych tvarov slov (v mil.)
prim-2.0	2005	250			
prim-2.1	2006	300	229	1,921	2,899
prim-3.0	2007	350	264	2,140	3,215
prim-4.0	2009	526	410	2,468	3,970
prim-5.0	2011	719	573	2,711	4,376
prim-6.0	1/2013	1 150	881	4,311	6,279
prim-6.1	9/2013	829	656	2,939	4,715

Tabuľka č. 2: Verzie a subkorpusy Slovenského hovoreného korpusu a ich rozsahy

Názov korpusu/subkorpusu		Prístupný od	Počet nahrávok	Časový rozsah	Počet tokenov
s-hovor-1.0		2008	71	12 hod., 45 min.	127 714
s-hovor-2.0		2010	154	72 hod., 17 min.	678 592
s-hovor-3.0		2011	246	180 hod., 23 min.	1 643 118
s-hovor-4.0		2012	353	282 hod., 37 min.	2 611 504
	s-hovor-4.0-sane	2012	291	154 hod., 40 min.	1 564 260
	s-hovor-4.0-upn	2012	62	127 hod., 57 min	1 047 244

Hlavnou zásadou budovania primárneho SNK i všetkých ďalších korpusov a databáz v komplexe projektov SNK je, že nová verzia je rozšírením predchádzajúcej, „pohlcuje“ ju a na vyhľadávanie jazykových informácií sa sprístupňuje záujemcom po získaní a spracovaní relevantného množstva dát a/alebo pri relevantnom skvalitnení anotácie. Prvé verzie (*prim0.1*, *prim0.2*) boli sprístupnené v rozpätí niekoľkých mesiacov, čo súviselo s testovacím režimom. Po dosiahnutí väčšieho množstva dát sa nové verzie hlavného korpusu od *prim-2.0* aktualizovali každý rok, od verzie *3.0* je to pravidelne každý druhý rok. Výnimkou je kratší rozostup medzi verziami *6.1* a *6.0* Vo verzii *prim-6.0* sa nachádzalo značné množstvo nesprávne skonvertovaných textov, čo bolo spôsobené zmenou grafického systému vo dvoch vydavateľstvách, takže bolo potrebné jej nahradenie opravenou verziou *6.1* V tomto prípade sa rozsah novej verzie korpusu nezväčšil, ale zmenšil.

Texty zaradené do korpusu sú podrobne bibliograficky a štýlovo-žánrovo anotované, všetky slovné tvary sú lematizované a morfologicky anotované – majú pri sebe informáciu o základnom tvare (leme) a gramatických vlastnostiach v danom kontexte.

Z vlastného veľkého korpusu sa od verzie *2.1* robia podkorpusy nielen podľa práva na prácu s textami a jazykovými informáciami prostredníctvom siete internet

alebo iba v rámci JÚĽŠ SAV, ukotveného v licenčnej zmluve, ale aj podľa hlavných štýlov a ďalších kritérií. Z korpusu obsahujúceho úplne všetky texty (*snk-all*, resp. od verzie *6.0 juls-all*) sa vyčlenia tie, ktoré môžu byť zaradené do korpusu prístupného na vyhľadávanie prostredníctvom siete internet, a vytvorí sa korpus *public-all*. Z neho sa potom tvoria všetky verejne prístupné podkorpusy príslušnej verzie:

- *-sane* – bez textov s nesprávnou diakritikou, bez textov spred roka 1955, z oblastí mimo Slovenska a z lingvistických časopisov,
- *-inf* – podkorpus publicistických (informatívnych) textov,
- *-prf* – podkorpus vedeckých, odborných a populárno-náučných textov,
- *-img* – podkorpus umeleckých textov,
- *-vyv* – vyvážený podkorpus (po jednej tretine z každého z uvedených štýlov),
- *-public-sk* – podkorpus všetkých pôvodných slovenských textov,
- *-img-sk* – podkorpus pôvodných slovenských umeleckých textov,
- *r55az89* – osobitný podkorpus textov z rokov 1955 – 1989.

Súčasťou každej novej verzie je celá škála informácií o štruktúre daného korpusu (zoznam bibliografických údajov všetkých textov a osobitne prekladov, zastúpenie textov podľa pôvodného jazyka diela, typu, žánru, vecnej oblasti) a o štatistických parametroch jednotlivých podkorpusov, resp. o jazykových a textových jednotkách, ktoré sú v nich obsiahnuté (frekvencie lem, slov, bigramov, trigramov, tetragramov slov, jazykové modely).

Okrem databáz *prim* a *s-hovor* sú súčasťou komplexu SNK ďalšie špecializované korpusy a databázy. Uvádzame tu iba aktuálne verzie a ich rozsahy, podrobnejšie o štruktúre a spôsobe tvorby niektorých z nich porov. napr. R. Garabík – L. Dimitrova (2012), M. Šimková – R. Garabík (2012; 2013):

- ručne morfologicky anotovaný korpus – *r-mak-4.0* (1,2 mil. tokenov),
- webový korpus – *sk-web-2.0* (1,05 mld. tokenov),
- paralelné korpusy:
  - slovensko-francúzsky (600 tis. tokenov),
  - slovensko-ruský (5,5 mil. tokenov),
  - slovensko-latinský (1,5 mil. tokenov),
  - slovensko-český (240 mil. tokenov),
  - slovensko-anglický (370 mil. tokenov),
- Slovenská terminologická databáza – 23 oblastí (približne 6 000 terminologických záznamov),
- slovenský WordNet – 25 tis. synsetov,
- litovský WordNet – 15 tis. synsetov,
- morfologická databáza – takmer 100 tis. slov a 3,3 mil. tvarov,
- Historický korpus slovenčiny – 370 tis. tokenov,
- pripravuje sa Korpus slovenských nárečí.

Celkovo sa v uvedených textových a jazykových materiálových zdrojoch nachádzajú aktuálne takmer 3 miliardy slovných a iných textových jednotiek. Časť korpusov (*prim-6.0-public-all, s-hovor-4.0, legal-1.1, web-1.1, web-1.2* po odstránení duplicitných textov a s malými úpravami v tokenizácii a lematizácii) je od r. 2013 prístupná verejnosti v spojenom súbore *Omnia*, ktorý zo zdrojov SNK vytvoril V. Benko primárne pre potreby pracovníkov oddelenia súčasnej lexikológie a lexikografie JÚĽŠ SAV.

Osobitnú súčasť práce SNK a verejnosťou najviac využívanú zložku predstavujú elektronické lingvistické, najmä lexikografické zdroje (http://slovniky.korpus.sk). Niektoré slovníky a databázy boli poskytnuté SNK na ďalšie spracovanie už v elektronickej podobe (Krátky slovník slovenského jazyka, Pravidlá slovenského pravopisu, Synonymický slovník slovenčiny a i.), mnoho tlačených diel sa však muselo často náročne digitalizovať a následne opravovať a spracúvať, celý proces si vyžadoval premyslenú logistiku, administráciu a v neposlednom rade aj finančnú investíciu. V špecifických prípadoch bolo treba hľadať riešenia netriviálnych problémov nielen na úrovni počítačového spracovania, ale aj z lingvistického hľadiska (porov. Garabík – Kajanová, 2012). Do lingvistických zdrojov sa okrem spomínaných slovníkov postupne dopĺňali napríklad tieto publikácie a celé vydania časopisov:

- A. Bernolák: Slowár Slowenskí Češko-Laťinsko-Ňemecko-Uherskí, 1825;
- Ľ. Štúr: Nauka reči Slovenskej, 1846;
- Ľ. Štúr: Nárečja Slovenskuo alebo Potreba písaňja v tomto nárečí, 1846;
- S. Czambel: Rukoväť spisovnej reči slovenskej, 1902;
- Pravidlá slovenského pravopisu s abecedným pravopisným slovníkom, 1931;
- Pravidlá slovenského pravopisu s pravopisným slovníkom, 1940;
- Morfológia slovenského jazyka, 1966;
- Dynamika slovnej zásoby súčasnej slovenčiny, 1989;
- Pramene k dejinám slovenčiny 1, 1992;
- Jazykovedný časopis, od r. 1954;
- Kultúra slova, od r. 1967;
- Československý terminologický časopis, 1962 – 1966;
- Slovenská reč, od r. 1932;
- Sociolinguistica Slovaca 1 – 6;
- Varia I – VIII;
- Slovenskí jazykovedci, 1925 – 1975.

Vďaka sprístupneniu lingvistickej produkcie JÚĽŠ SAV a ďalších relevantných titulov na internete sú výsledky slovenskej lingvistiky dostupné komukoľvek, kdekoľvek a kedykoľvek, čo oceňujú nielen slovakisti v zahraničí, ale aj lingvisti a iní záujemcovia na území Slovenska.

Stúpajúca kvantita i kvalita výsledkov práce kolektívu SNK a jedinečnosť tohto pracoviska, ktoré disponuje potrebnými elektronickými materiálovými zdrojmi a nástrojmi na výskum a počítačové spracovanie slovenčiny, prispeli k rozvinutiu spolupráce v mnohých domácich i zahraničných projektoch, vďaka ktorým sa mohli jednotlivé databázy lepšie využívať, zároveň aj skvalitňovať a rozširovať (uvádzame ich podľa roku začatia):

- ***Využitie spoločných vlastností češtiny a slovenčiny na budovanie anotovaných národných jazykových korpusov*** – Ústav formálnej a aplikovanej lingvistiky MFF UK Praha, 2004 – 2005.
- ***Morfosyntaktická analýza Slovenského národného korpusu***, VEGA – FF PU Prešov, 2004 – 2006.
- ***Budovanie paralelných korpusov (slovensko-chorvátsky a slovensko-ruský korpus)***, VEGA, 2005 – 2007.
- ***MONDILEX***, FP7 – Bulharsko, Poľsko, Slovinsko, Rusko, Ukrajina, 2008 – 2010
- ***Konfrontačný výskum kolokácií v slovenčine a v nemčine***, VEGA – FF UCM Trnava, 2008 – 2010.
- ***Slovak Online*** – Nemecko, Poľsko, Litva, 2009 – 2011.
- ***Spracovanie obchodnovednej terminológie pre potreby Slovenskej terminologickej databázy s dôrazom na analýzu terminologických neologizmov***, VEGA – Ekonomická univerzita Bratislava, 2009 – 2011.
- ***Svedkovia obdobia útlaku*** – Ústav pamäti národa, 2010 – 2011.
- ***EuroMatrixPlus***, FP7 – Nemecko, V. Británia, ČR, USA, Taliansko, Francúzsko, Írsko, Bulharsko, 2010 – 2012.
- ***CESAR*** – Maďarsko, Chorvátsko, Poľsko, Srbsko, Bulharsko, 2011 – 2013.
- ***Konfrontační popis současného českého a slovenského lexika (systémové vztahy a komunikační koexistence)***, GAČR – ÚJČ AV ČR Praha, 2011 – 2014.
- Sémantická a distribučná analýza adjektív v nemčine a slovenčine, VEGA – FF UCM Trnava, 2011 – 2013.
- ***Sila svedectva*** – Ústav pamäti národa, 2011 – 2012.
- ***NetWords*** – 16 krajín, 2011 – 2015.
- ***Hľadanie pravdy*** – Ústav pamäti národa, 2012 – 2014.
- ***lingvo.info*** – Belgicko, Nemecko, Poľsko, Dánsko, Slovinsko, Litva, 2012 – 2014.
- ***COST/PARSEME*** – 27 krajín, jún 2013 – 2017.

Jedným z významných výsledkov zahraničnej spolupráce je publikácia The Slovak Language in the Digital Age – Slovenský jazyk v digitálnom veku (Šimková a kol., 2012), pripravená s podporou projektu CESAR a METANET v spolupráci

s autormi z celého Slovenska v rámci Série bielych kníh o stave jazykov a jazykových technológií v Európe. Z hodnotenia úrovne podpory jazykových technológií štátnymi inštitúciami alebo komerčnými firmami vyplynulo, že 21 z 30 analyzovaných jazykov nemá žiadnu alebo má len slabú podporu a ich digitalizácia je nedostatočná. Medzi ne sa, napriek všetkým vybudovaným zdrojom, radí aj slovenčina, v ktorej je veľmi nedostatočne rozvinutá oblasť strojového prekladu, chýba sémantický korpus a viaceré špecializované nástroje na spracovanie písanej a hovorenej podoby slovenčiny. Cieľom Série bielych kníh bolo upozorniť vlády a komerčné sféry krajín so slabou podporou vývoja jazykových technológií na zaostávanie v tejto oblasti a iniciovať ich záujem o zlepšenie situácie najmä pre malé jazyky.

Súčasťou zapojenia sa SNK do medzinárodného kontextu sú bienálne konferencie SLOVKO, ktoré kolektív oddelenia SNK organizuje od r. 2005, keď prevzalo štafetu od organizátorov prvých dvoch ročníkov A. Jarošovej a V. Benka. Z každej konferencie (okrem druhého ročníka) vyšiel súbor príspevkov, od. r. 2007 ho účastníci dostávajú už pred začatím konferencie:

- Computer Treatment of Slavic and East European Languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2005. Ed. R. Garabík. Bratislava: Veda 2005. 246 s.
- Computer Treatment of Slavic and East European Languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2007. Ed. J. Levická – R. Garabík. Brno: Tribun 2007. 318 s.
- NLP, Corpus Linguistics, Corpus Based Grammar Research. Zborník z medzinárodnej vedeckej konferencie Slovko 2009. Ed. J. Levická – R. Garabík. Brno: Tribun 2009. 401 s.
- Natural Language Processing, Multilinguality. Zborník z medzinárodnej vedeckej konferencie Slovko 2011. Ed. D. Majchráková – R. Garabík. Brno: Tribun 2011. 180 s.
- Natural Language Processing, Corpus Linguistics, E-Learning. Zborník z medzinárodnej vedeckej konferencie Slovko 2013. Ed. K. Gajdošová – A. Žáková. Lüdenscheid: RAM-Verlag 2013. 303 s.

Od začiatku fungovania SNK sa veľká pozornosť venovala nielen poskytovateľom textov, bez ktorých by nemohol byť žiaden korpus (od r. 2002 bolo oslovených vyše 1 100 právnických i fyzických osôb, zmluvy sú uzavreté s takmer 700 z nich), ale aj všetkým používateľom – záujemcom o poznávanie jazykového systému a fungovania jazykových prostriedkov prostredníctvom elektronických zdrojov. Už v r. 2003 uskutočnili pracovníci SNK (M. Šimková, R. Garabík, A. Horák) prvé prednášky a workshopy o možnostiach využitia korpusov v lingvistickej práci pre kolegov a študentov Filozofickej fakulty Prešovskej univerzity v Prešove, kde je korpus

stále veľmi využívaným zdrojom bádania. Rady používateľov sa od Prešova rozšírili do Košíc, z bratislavských slovakistických a počítačových pracovísk (JÚĽŠ SAV, PdF UK, MFF UK, FIIT STU) aj do Trnavy a Nitry. Možnosť práce s korpusovým materiálom prostredníctvom siete internet otvorila priestor na intenzívnejší výskum slovenčiny aj zahraničným bádateľom – stálymi registrovanými používateľmi sú lingvisti z ČR, Poľska a ďalších slovanských krajín, ale aj slovakisti, slavisti, ba i nelingvisti z Nemecka, Rakúska, Švajčiarska, Holandska, Francúzska, Nórska, USA, Číny a ďalších krajín. Značnú časť všetkých registrovaných používateľov, ktorých počet sa postupne zvýšil na vyše 700 ročne, tvoria v súčasnosti učitelia základných a stredných škôl. Okrem stále pokračujúcich nepravidelných prezentácií, prednášok a workshopov sa od r. 2013 konajú na pôde SNK pod vedením K. Gajdošovej pravidelné semináre pre začiatočníkov i pokročilejších používateľov korpusových databáz (aktuálne informácie sú vždy na stránke korpus.sk, podrobnejšie porov. aj Gajdošová, 2013). Od mája do septembra 2013 sa na seminároch zúčastnilo takmer 150 záujemcov, z toho 55 priamo v SNK JÚĽŠ SAV.

Keď J. Mistrík v Encyklopédii jazykovedy konštatoval, že „Náš jazyk stojí dnes na úrovni najmodernejších jazykov sveta“ (c. d.), mal zrejme na mysli jazyk ako taký spolu s vtedajšími poznatkami o ňom. Nazdávame sa, že toto vyjadrenie, napriek vedomiu nedostatočného pokrytia všetkých oblastí jazyka a napriek viacerým chýbajúcim zdrojom a nástrojom, môžeme dnes rozšíriť aj o stav elektronických zdrojov, počítačového spracovania slovenčiny a korpusovej lingvistiky na Slovensku, k čomu prispela aj istá nevyhnutná súhra vonkajších a vnútorných faktorov. Vďaka existencii a výsledkom Slovenského národného korpusu sa podarilo nielen prekonať veľké zaostávanie Slovenska v tejto oblasti, ale SNK sa stalo špičkovým a veľmi efektívne pracujúcim kolektívom, ktorý rozsahom a obsahom práce dosahuje zaradenie medzi medzinárodne uznávané korpusové pracoviská, hoci tie disponujú podstatne väčším finančným rozpočtom aj počtom zamestnancov. V databázach SNK sa v súčasnosti nachádzajú rozsiahle materiálové zdroje s množstvom jazykových informácií a stále sa zlepšujúcimi nástrojmi na ich vyhľadávanie, selektovanie a usporadúvanie, ktoré umožňujú spracovať rozmanité výskumné úlohy. Neporovnateľne s doterajšími možnosťami jazykovedného výskumu i ďalších vedných odborov (literárna veda, história, psychológia, filozofia, etnológia, logopédia, informatika) sa v týchto zdrojoch dajú sledovať stopy minulého vývinu nášho jazyka, obraz sveta uplynulých desaťročí, ako aj predikovať ďalšie procesy. Elektronicky spracované a archivované zdroje slúžia zároveň na uchovávanie kultúrneho dedičstva pre budúce generácie.

Za 20 rokov od vydania Encyklopédie jazykovedy a za 11 od vzniku oddelenia Slovenského národného korpusu JÚĽŠ SAV sa situácia v existencii a poznaní korpusu významne zmenila. Heslá ***Slovenský národný korpus, korpusová lingvistika, po-***

čítačové spracovanie prirodzeného jazyka sa chystajú na zaradenie do príslušných zväzkov Encyklopédie Beliana, lexémy ***korpus, korpusový*** sa nachádzajú aj vo všeobecnom výkladovom Slovníku súčasného slovenského jazyka H – L (2011), kde sa už korpus nevykladá ako „ohraničený súbor jazykových výpovedí zaznamenaných písomne, príp. na magnetofónovej páske al. na platni", ale ako ***„rozsiahly súbor textov v elektronickej podobe"***. Napĺňaním leitmotívu tvorby korpusov „čím viac dát, tým sú to lepšie dáta" sa v súčasnosti dospieva k natoľko veľkým rozsahom korpusov (miliardy jednotiek), že ich ohraničenosť sa relativizuje a v skutočnosti sa sleduje cieľ obsiahnuť také množstvo jazykových javov, ktoré by dostatočne reprezentovali alebo dokladali celý systém daného prirodzeného jazyka, resp. sa blížili až k hranici obsiahnutia všetkých jazykových javov. Tento rozmer predpokladá nielen kvantitatívny nárast textov, ale aj presnejšiu a všestrannejšiu lingvistickú anotáciu – v korpuse sa dajú nájsť len také prostriedky a také jazykové informácie, aké doň vložíme. V tomto smere čaká Slovenský národný korpus a korpusovú lingvistiku ešte veľa práce.

Literatúra

Encyklopédia jazykovedy. Zost. J. Mistrík. Bratislava: Obzor 1993. 513 s.

GAJDOŠOVÁ, Katarína: Cyklus prezentácií a praktických seminárov Vyhľadávanie v Slovenskom národnom korpuse. In: Slovenská reč, 2013, roč. 78, č. 3 – 4, s. 209 – 211.

GAJDOŠOVÁ, Katarína – ŠIMKOVÁ, Mária: Slovenský hovorený korpus (2008 – 2012). In: Jazykovedné štúdie XXXI. Ed.: K. Gajdošová – A. Žáková. Bratislava: Veda 2013 (v tlači).

GARABÍK, Radovan – DIMITROVA, Ludmila: Bilingual Corpus – Digital Repository for Preservation of Language Heritage. In: Digital Presentation and Preservation of Cultural and Scientific Heritage. Museology & Heritage Studies II/2012, s. 132 – 141.

GARABÍK, Radovan – KAJANOVÁ, Michaela: Problémy a výsledky počítačového spracovania diela *Slowár Slowenski Češko-Laťinsko-Ňemecko-Uherski seu Lexicon Slavicum Bohemico-Latino-Germanico-Ungaricum*. In: Slovo v slovníku. Ed. K. Buzássyová – B. Chocholová – N. Janočková. Bratislava: Veda 2012, s. 294 – 300.

Insight into the Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2006. 208 s.

JAROŠOVÁ, Alexandra: Malá inventúra pred hľadaním spoločného jazyka. In: Slovenčina a čeština v počítačovom spracovaní. Ed. A. Jarošová. Bratislava: Veda 2001, s. 7 – 10.

LEECH, Geoffrey: Corpus Annotation Schemes. In: Literary and Linguistic Computing, 1993, Vol. 8, No. 4, s. 275 – 281. Cit. podľa Leech, Geoffrey: Anotační systémy pro značkování korpusů. In: Acta Universitatis Carolinae. Philologica 3 – 4. Studie z korpusové lingvistiky. Praha: Univerzita Karlova – Nakladatelství Karolinum, s. 185 – 197.

PÁLEŠ, Emil: SAPFO. Parafrázovač slovenčiny. Počítačový nástroj na modelovanie v jazykovede. Bratislava: Veda 1994. 308 s.

Projekt budovania Národného korpusu slovenského jazyka a projekt elektronizácie jazykovedného výskumu v rokoch 2002 – 2006. Materiál predložený na rokovanie vlády SR. Dostupný na WWW: http://www.rokovania.sk

Slovník súčasného slovenského jazyka. H – L. Hl. red. K. Buzássyová – A. Jarošová. Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied 2006. 1134 s.

ŠIMKOVÁ, Mária: Slovak National Corpus – History and Current Situation. In: Insight into Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2006, s. 152 – 159.
ŠIMKOVÁ, Mária: Korpusová lingvistika na Slovensku. In: Jazykovedný časopis, 2008, roč. 59, č. 1 – 2, s. 11 – 24.
ŠIMKOVÁ, Mária: Язык – корпус – словарь. In: 70 години българска академична лексикография. Sofia: Akademično izdatelstvo „Prof. Marin Drinov“ 2013, s. 39 – 47.
ŠIMKOVÁ, Mária – GARABÍK, Radovan: The Slovak National Corpus and its Corpus Linguistic Resources. In: Prace filologiczne, tom LXIII. Warszawa: Wydział polonistyki Uniwersytetu Warszawskiego 2012, s. 109 – 119.
ŠIMKOVÁ, Mária a kol.: The Slovak Language in the Digital Age – Slovenský jazyk v digitálnom veku. White Paper Series/Séria bielych kníh. Ed. G. Rehm – H. Uszkoreit. Berlin – New York: Springer 2012. 85 s. (dostupná na WWW: http://www.meta-net.eu/whitepapers/volumes/slovak)
ŠIMKOVÁ, Mária – GARABÍK, Radovan: Slovenský národný korpus (2002 – 2012):
východiská, ciele a výsledky pre výskum a prax. In: Jazykovedné štúdie XXXI. Ed.: K. Gajdošová – A. Žáková. Bratislava: Veda 2013 (v tlači).