JAZYKOVEDNÝ ČASOPIS, 59, 2008, 1 – 2

KORPUSOVÁ LINGVISTIKA NA SLOVENSKU

MÁRIA ŠIMKOVÁ

ŠIMKOVÁ, Mária: Corpus Linguistics in the Slovak Republic. Jazykovedný časopis, 2008, Vol. 59, No. 1 – 2, pp. 11 – 24. (Bratislava)

Slovakia joined the worldwide trend of the development of computer and linguistic technologies as late as 1990. The first corpus of texts of the Slovak language was gradually made available up to 2002 for internal use within the framework of the Ľ. Štúr Institute of Linguistic of the Slovak Academy of Sciences in Bratislava. The new work on the building of the Slovak National Corpus in the same place starting in 2003 started with the preparation of a concept of the structure of data in the corpus and methods for their primary processing, i. e., conversion, tokenization, bibliographic and style-genre annotation. Later there was prepared the morphological tagset proceeded with manual morphological annotation. The first collective systematic work focused on obtaining extensive textual material and its consequent linguistic processing, prepared and provided by the Slovak National Corpus, was the Morphosyntactic Analysis of the Slovak National Corpus Project in the partnership with University of Prešov. In its current form, the SNC provides the basic research material for all categories of users and anybody who is interested in the Slovak language – for the creation of new dictionaries and databases, like the Slovak Terminological Database, for updating orthographic or grammar manuals and for the specialists in the NLP.

1. Lingvistické bádania v súčasnosti charakterizuje vysoká dynamika, interdisciplinarita, rozmanitosť škôl a smerov. Jednou z moderných a veľmi dynamicky sa rozvíjajúcich oblastí je korpusová lingvistika ako hlavná zložka aplikovanej počítačovej (komputačnej) lingvistiky (Čermák, 2000), ktorá spolu s teoretickou kvantitatívnou (štatistickou) a algebrickou lingvistikou tvorí odbor matematickej lingvistiky (Černý, 1996). Korpusová lingvistika skúma a opisuje jazyk, resp. jeho použitie (parole) v reálnych kontextoch, ktoré sa vyhľadávajú v korpuse a spracúvajú pomocou počítačových nástrojov a štatistických metód. Od predchádzajúceho obdobia lingvistických výskumov materiálovo založených na individuálnej excerpcii, prípadne na všeobecnejších, no predsa ťažšie dostupných kartotékach, sa práca s korpusom odlišuje viacerými parametrami a metódami:

a) takmer výlučné postavenie jazykovej introspekcie a intuície sa oslabuje možnosťami exaktnej analýzy dostatočného množstva dostatočne rôznorodého materiálu;

b) nebývalá rozsiahlosť materiálu kladie zvýšené nároky na selekciu relevantných kontextov a javov, a teda aj na čas;

c) lingvistika už nevystačí len s perom a papierom, potrebuje náležité počítačové zabezpečenie a lingvista musí disponovať značnými počítačovými zručnosťami;

d) rovnaký materiál a možnosti jeho skúmania sú v princípe k dispozícii kedykoľvek a komukoľvek;
e) výsledky výskumu sú opakovane verifikovateľné;
f) korpusový lingvista, resp. korpusová lingvistika rieši otázky, s ktorými sa predtým nebolo potrebné vyrovnávať v takej perspektíve, napr. rozsah a reprezentatívnosť materiálovej bázy, stálosť a súčasne aktuálnosť materiálovej bázy; nejazykové súčasti textu, formálna a štylistická chybovosť textov (myslia sa tým skutočné preklepy, pravopisné chyby a štylizačné lapsusy, ktoré sa vyskytujú aj v korigovaných a oficiálne vydaných textoch); formalizácia gramatických opisov, premenlivosť gramatických teórií a pravopisných či ortoepických kodifikácií; nespoľahlivosť automatizovaných procedúr.

Nevyhnutným predpokladom vzniku a rozvoja korpusovej lingvistiky je existencia korpusu – jazykovej databázy písaných a/alebo hovorených textov, ktoré sú počítačovo uložené, zvyčajne verejne prístupné, v súčasnosti už štandardne lingvisticky predspracované (lematizované a aspoň morfologicky anotované) a cielene štruktúrované ako vyvážený alebo reprezentatívny celok. Negativisticky orientovaná otázka „Načo korpus?“, formulovaná niektorými lingvistami, ale aj nelingvistami ešte pred dvoma či troma desiatkami rokov (porov. Čermák, 2001 a i.) a odrážajúca neistotu z nástupu nových metód výskumu, sa dnes zmenila na otázku „Ako korpus čo najlepšie využiť?“. Rovnako sa prekonal druhý extrém zvyčajne sprevádzajúci objavenie nových možností výskumu a začiatky ich využívania – privysoké, nereálne očakávania či až absolutizácia korpusovej databázy ako kategorického imperatívu pri akomkoľvek jazykovednom výskume, ako aj pri rozhodovaní sa o gramatickosti či negramatickosti jazykovej jednotky: „... dostatečně velký korpus (sbírka dokladů parole) sice může být „objektivizací“ této parole, avšak z principiálních důvodů nemůže být „objektivizací“ langue... základním materiálem pro gramatický výzkum (tj. pro explikaci langue) je právě tato langue (jako implicitní znalost jazyka), korpus může být pro takovou práci „pouze“ inspirací a korektivem“ (Oliva – Doležalová, 2004, s. 10). Je totiž otázne, či sa môže podariť vytvoriť taký korpus živého prirodzeného jazyka, v ktorom by boli adekvátne zastúpené všetky jeho jazykové prostriedky tak, aby bolo možné povedať, že to, čo nie je v korpuse, nie je ani v jazyku. Ako ukazujú autori citovanej štúdie, v jazyku sa nachádzajú (a v parole sa aj určite používajú) prostriedky, ktoré sa v korpuse, hoci je dosť veľký, nedajú nájsť, a naopak, korpus obsahuje konštrukcie i jednotlivé jednotky, ktoré skutočne nie sú v súlade s jazykovým systémom a nejde pritom o preklepy či varianty (napr. vokatív privlastňovacieho zámena *tvůj*, čo je v češtine teoreticky vylúčený tvar; c. d., s. 7).

Lingvisticky je teda korpus zdrojom informácií – a dobre koncipovaný a anotovaný korpus môže byť zdrojom veľmi dobrých, bez korpusu často ani nezískateľných informácií – a korpusová lingvistika je predovšetkým súhrnom metód na získanie potrebných informácií z jazykového korpusu. Základný postup a možnosť sú napokon rovnaké ako všade inde: od jazykovej formy cez jej posúdenie v kontexte sa prichádza

k zovšeobecneniu jej vlastností, významu a funkcie (porov. Čermák, 2006, s. 12), je však k dispozícii neporovnateľne rozsiahlejší materiál a nástroje na jeho využitie. Korpusová lingvistika vo svete sa v súčasnosti môže opierať o množstvo rôznych korpusov viacerých jazykov a má za sebou nemalé výsledky, ktoré predstavujú jazyk a jeho jednotky v novom svetle.

Najväčší rozsah, v súčasnosti stovky miliónov, ba aj miliardy textových jednotiek (tokenov), majú všeobecné, národné korpusy jednotlivých jazykov (sú to najmä európske jazyky a americká angličtina, ale existujú aj menšie korpusy iných jazykov, napr. taiwanský korpus, krymskotatársky korpus či korpus starej babylončiny), ktoré slúžia nielen na lingvistické výskumy vrátane tvorby slovníkov, ale aj pre prekladateľov, učiteľov, logopédov, neurológov, odborníkov na počítačové spracovanie prirodzeného jazyka a pod. V súčasnom trende globalizácie a informatizácie je osobitne pre malé jazyky veľmi dôležité, aby boli k dispozícii ich kvantitatívne čo najrozsiahlejšie databázy s kvalitatívne čo najpodrobnejšou a najpresnejšou anotáciou. Výhodou je aj existencia špecializovaných korpusov a databáz, napr. korpusu hovorených prejavov, paralelných korpusov jedného jazyka s iným jazykom alebo viacerými inými jazykmi, ale aj terminologických databáz a elektronických lingvistických zdrojov, najmä slovníkov.

2. Na Slovensku sa základy matematickej lingvistiky začali budovať v 60. rokoch 20. storočia, najmä v oblasti teoretickej algebrickej lingvistiky zásluhou J. Horeckého (1969). Kvantitatívnej zložke tohto odboru sa vtedy u nás aktívne venoval aj G. Altmann (1963; 1969; 1972), neskôr J. Mistrík (1969, 1976, 1985). Prvý elektronický korpus textov slovenského jazyka sa začal tvoriť v Jazykovednom ústave Ľudovíta Štúra Slovenskej akadémie vied v Bratislave až začiatkom 90. rokov 20. storočia (o histórii matematickej lingvistiky na Slovensku porov. aj Šimková, 2004). Po príprave koncepcie korpusu a lexikálnej bázy dát (Horecký, 1990) postupne vznikal v JÚĽŠ SAV v rámci minimálnych finančných i personálnych možností malý interný korpus textov slovenského jazyka. V závere jeho budovania v r. 2001 obsahoval tento korpus necelých 30 miliónov textových jednotiek vrátane lexikografických príručiek a predstavoval jednoduchú lexikálnu databázu bez akejkoľvek lingvistickej anotácie. V tejto podobe slúžil korpus na pôde JÚĽŠ SAV najmä lexikografom pri koncipovaní nového výkladového slovníka slovenčiny a pri aktualizácii nových vydaní existujúcich lexikografických príručiek.

Medzinárodne porovnateľný jazykový korpus sa na Slovensku začal systematicky budovať po inštitucionalizovaní oddelenia Slovenského národného korpusu v JÚĽŠ SAV v r. 2002 na základe osobitnej dohody medzi Ministerstvom školstva SR, Ministerstvom kultúry SR a Slovenskou akadémiou vied. Otvorili sa tak nové možnosti pre lingvistický výskum súčasného slovenského jazyka, ako aj pre rovnoprávne postavenie slovenčiny v pluralitnom celku európskych jazykov v období informačnej spoločnosti. Primárnou úlohou bola tvorba Slovenského národného korpusu ako všeobecnej elektronickej databázy súčasnej slovenčiny (http://korpus.juls.savba.sk), ktorej dolná hranica sa stanovila na rok 1955 v súvise s pravopisnou reformou v r. 1953

a jej zavedením do praxe, horná hranica bola pôvodne stanovená na rok 2005 v súvise s dĺžkou trvania schváleného projektu. Aktuálne je horná hranica posunutá do roku 2011, dokedy bola schválená druhá etapa projektu Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku.

Pomenovanie Slovenský národný korpus zahŕňa nielen vlastnú, primárnu databázu veľkého národného korpusu, ale aj viacero čiastkových úloh, ktoré spolu súvisia a výsledkom ich riešenia je celok skladajúci sa z niekoľkých, do istej miery samostatných zložiek. V prvej fáze mal projekt splniť tieto ciele: zachytiť jazyk v celej jeho šírke (novinové texty, beletria, odborné publikácie, hovorený jazyk a pod.) na základe lingvisticky zdôvodnených kritérií a tým vytvoriť objektívny a autentický zdroj jazykovej informácie, ktorý by bol materiálovým východiskom na všestranný jazykovedný výskum, tvorbu základných akademických diel (viaczväzkového slovníka súčasnej slovenčiny, ortoepického slovníka, frekvenčného a retrográdneho slovníka, akademickej gramatiky a lexikológie slovenského jazyka), ako aj na aktualizáciu jestvujúcich praktických jazykových príručiek (Krátkeho slovníka slovenského jazyka, pravopisného slovníka v Pravidlách slovenského pravopisu, Synonymického slovníka slovenčiny). Súčasťou Slovenského národného korpusu mali byť osobitné korpusy nárečových, historických a hovorených textov a paralelné korpusy. Vzhľadom na závažnosť legislatívnej sféry sa počítalo s vytvorením špecializovaného podkorpusu – databázy legislatívnych textov – s cieľom podporovať zvyšovanie terminologickej kultúry v oblasti právneho jazyka v spolupráci so zainteresovanými vládnymi orgánmi (napr. legislatívnou radou vlády, pracoviskom Jednotného automatizovaného systému právnych informácií Ministerstva spravodlivosti SR a Ústavom pre aproximáciu práva).

Rozsiahlosť stanovených úloh odrážala potreby slovenskej lingvistiky a odborníkov na počítačové spracovanie prirodzeného jazyka (natural language processing – NLP), ktorých napĺňanie zaostávalo za svetovým trendom rozvoja informačných a jazykových technológií minimálne jedno až dve desaťročia. (Prvé elektronické korpusy boli vytvorené v 60. rokoch 20. st. na materiáli anglického jazyka, v 90. rokoch nastal obrovský rozmach korpusov a korpusovej lingvistiky v celej Európe.) V závere prvej etapy projektu sa ukázalo, že osemčlenný pracovný tím nemôže splniť všetky úlohy, ktoré napr. v Českej republike rieši približne sto odborných a vedeckých pracovníkov na štyroch špecializovaných pracoviskách v Prahe a Brne. Niektoré úlohy sa čiastočne modifikovali, iné boli v plnom rozsahu presunuté do druhej etapy (napr. hovorený korpus) alebo úplne vylúčené z úloh Slovenského národného korpusu v najbližšom časovom horizonte (nárečový korpus a historický korpus, ktoré vyžadujú účasť odborníkov z príslušných oblastí).

Začiatok budovania Slovenského národného korpusu (ďalej SNK) v r. 2002 predstavovalo fyzické vybudovanie nových priestorov v podkroví budovy Jazykovedného ústavu Ľ. Štúra SAV v Bratislave a ich vybavenie potrebnou počítačovou technikou. Riešiteľský kolektív sa takisto budoval postupne z externého prostredia – JÚĽŠ SAV ani žiadna vysoká škola na Slovensku nepripravovala a nepripravuje

špecialistov na korpusovú lingvistiku. Na pôde nového oddelenia preto prebiehali semináre z korpusovej a počítačovej lingvistiky, na ktorých odznelo vyše 30 prednášok, z toho 20 zahraničných (časť z nich je zhrnutá v zborníku *Insight into the Slovak and Czech Corpus Linguistics*, 2006). Významnú pomoc poskytli najmä české počítačovo- a korpusovolingvistické pracoviská v Prahe i Brne, na ktorých jednotliví pracovníci SNK JÚĽŠ SAV absolvovali aj krátkodobé študijné pobyty. Na základe získaných poznatkov a podkladov sa vypracovala rámcová koncepcia budovania SNK, komplexného počítačového spracovania slovenského jazyka a elektronizácie jazykovedného výskumu na Slovensku, v ktorej bolo primárne stanovenie cieľa a rekognoskácia potenciálnych používateľov korpusu a ich potrieb (Šimková, 2003; 2004a). Súčasťou koncepcie bola stratifikácia textov v Slovenskom národnom korpuse na základe sociolingvistickej analýzy štýlovo-žánrovej distribúcie textov súčasného slovenského jazyka opretej o sondážne prieskumy spojené s prípravou viaczväzkového výkladového slovníka slovenského jazyka, ktorý sa koncipuje v JÚĽŠ SAV, a o podobné výskumy a skúsenosti Ústavu Českého národného korpusu (porov. napr. Králík, 2004; Šimková, 2004b).

Nevyhnutnou súčasťou prípravy budovania elektronického jazykového korpusu sú viaceré základné dokumenty: licenčná zmluva na poskytnutie textov na nekomerčné použitie podľa autorského zákona (umožniť SNK získať automaticky text na základe osobitného paragrafu v autorskom zákone sa nepodarilo ani pri jeho novelizácii), podmienky používania korpusu, internetová stránka a pod. Postupne sa pripravovali zásady technického spracúvania textov: konverzie elektronických verzií na jeden formát, skenovanie a zhotovovanie technických rozmnožení starších textov, odstraňovanie grafických a iných netextových súčastí textov – tabuliek, obrázkov, šachových partií a pod. (Domin, 2004; Šmotlák – Hladík, 2004), zásady tokenizácie a lematizácie (segmentácie a prvotného lingvistického spracovania) získaných textov (Garabík, 2005), ako aj zásady bibliografickej a štýlovo-žánrovej anotácie (Garabík, 2004). Na základné lingvistické spracovanie textu (lematizáciu a morfologickú anotáciu) bol zapožičaný softvér z Ústavu formálnej a aplikovanej lingvistiky Matematicko-fyzikálnej fakulty Univerzity Karlovej v Prahe, na prácu s korpusom (na základe registrácie) sa zakúpil korpusový manažér Manatee s klientom Bonito z Fakulty informatiky Masarykovej univerzity v Brne. V súčasnosti sa interne používa aj Sketch Engine (Kilgarriff – Rychly – Smrz – Tugwell, 2004), ktorý okrem absolútnej frekvencie kolokácie automaticky normalizuje hodnoty pravdepodobnosti spoločného výskytu dvoch slov a získané zoznamy slov triedi aj podľa syntaktických funkcií.

Kým prvé verzie hlavného, primárneho korpusu (prim0.1, prim0.2, prim1, prim-2.0) obsahovali predovšetkým publicistiku a najmä texty z posledného desaťročia (ide o pomerne ľahko získateľné a spracovateľné dáta, keďže tieto texty sú zväčša dostupné priamo v elektronickej podobe), posledné verzie vyvážených korpusov už dosahujú aj lepší pomer štýlového zastúpenia (60 % publicistiky, po 20 % umeleckých a odborných textov), aj lepšie zastúpenie textov z obdobia rokov 1955

až 1992 (dokedy v podstate neexistovali elektronické verzie textov). Vzhľadom na to bolo možné a pre špecifické potreby niektorých výskumov priamo potrebné prikročiť od verzie prim-2.1 k vytváraniu samostatných podkorpusov s publicistickými textami, odbornými a populárno-náučnými textami a s umeleckými textami – spomedzi nich sa osobitne vyčleňujú originálne (pôvodné) slovenské umelecké texty. V pripravovanej verzii prim-4.0 sa už plánujú aj podkorpusy vytvorené podľa časového kritéria: od roku 1955 do roku 1989 a od roku 1990 do súčasnosti. Hranica bola zvolená v súvise so zmenou spoločensko-politických pomerov a výraznou dynamikou jazyka v nových podmienkach. Štýlovo-žánrové a časové vyvažovanie textov sa pri budovaní korpusov dosahuje postupne. V SNK na to bolo potrebných niekoľko rokov intenzívneho získavania jednotlivých textov v rámci autorského zákona (bolo oslovených vyše tisíc, ale zazmluvnených „len“ takmer päťsto poskytovateľov) a masívne skenovanie a upravovanie do elektronickej podoby tých textov, ktoré existujú iba v papierovej forme. V poslednom období sa v SNK naskenuje, rozpozná a rekonštruuje každoročne vyše 60 tisíc strán textov, čo je nemysliteľné bez externých spolupracovníkov, najmä študentov.

Okrem už spomínanej otázky o (ne)potrebe korpusov sa v prvých dvoch desaťročiach ich vytvárania vo svete diskutovalo aj o (ne)potrebe vnášania lingvistických informácií do lexikálnych databáz, ktoré slúžili predovšetkým na lexikografické ciele. Dôvodov na odmietanie lingvistických anotácií alebo aspoň na veľmi rezervovaný postoj k nim bolo viacero, niektoré sme formulovali v úvode (premenlivosť gramatických teórií a pod.), niektoré sa vynorili s pribúdaním korpusov flektívnych jazykov. Možnosti využitia lingvistických informácií v korpuse napriek ich nedokonalosti (napr. 95 % úspešnosť automatizovanej morfologickej anotácie sa považuje za vysokú, ale reálne to znamená, že každá dvadsiata značka je chybná, čo je značne nespoľahlivé) sa však stále zväčšujú a postupne prevýšili problémy a investície pri ich vkladaní do korpusu. V súčasnosti je lingvisticky anotovaný korpus štandardným typom korpusu a základným východiskom pre korpusovolingvistické analýzy daného jazyka.

V SNK sa už v koncepcii počítalo a aj sa pomerne rýchlo začalo s prípravou vlastného morfologického tagsetu (súboru pravidiel a značiek) pre prvú, morfologickú rovinu internej lingvistickej anotácie (Forróová – Garabík – Gianitsová – Horák – Šimková, 2003; Forróová – Horák, 2004; Garabík – Gianitsová – Horák – Šimková, 2004). Po prvých testovacích verziách korpusu bez lematizácie a morfologickej anotácie sa v r. 2004 sprístupnil korpus plne lematizovaný a morfologicky automatizovane označkovaný. Vzhľadom na to, že išlo o morfologickú anotáciu pomocou českého softvéru a s použitím značiek i princípov vytvorených pre český jazyk, táto verzia bola prístupná iba na interné použitie. Aj táto prvá podoba morfologickej anotácie, s vedomím rozdielov v jazykových systémoch i gramatických teóriách oboch jazykov a s obmedzeniami technicko-inžinierskeho východiska českého tagsetu, však poslúžila jednak na prvé analýzy v rámci spoločného výskumného projektu s Filozofickou fakultou Prešovskej univerzity v Prešove (porov. ďalej), jednak na predanotovanie textov, ktoré

sa anotovali ručne na základe vlastného slovenského tagsetu, aby sa vytvoril dostatočný trénovací a testovací súbor ručne anotovaných dát (Garabík – Gianitsová-Ološtiaková, 2005; Karčová – Šimková, 2006). Zároveň sa začalo s vývojom vlastného morfologického analyzátora a generátora tvarov slovenského jazyka (Garabík, 2005; 2007), aby sa čo najviac zlepšila automatizovaná morfologická anotácia. Slovník slovenského morfologického analyzátora v súčasnosti obsahuje všetky slová z Krátkeho slovníka slovenského jazyka (2003) doplnené o najfrekventovanejšie slová z korpusu, frekventované skratky, značky a vlastné mená: spracovaných je vyše 60 tisíc slov v rámci takmer 1 500 vzorov s úplnými paradigmami.

Ručne morfologicky anotované podkorpusy r-mak boli a sú zverejňované spolu s novými verziami všeobecného, základného korpusu prim od r. 2006. Základný korpus narastá v posledných rokoch pravidelne o vyše 50 miliónov textových jednotiek ročne, ručne morfologicky anotované podkorpusy sa v ročných časových intervaloch približne zdvojnásobovali: od prvého rozsahu 322 600 textových jednotiek v r-mak-1.0 cez 512 tisíc textových jednotiek v r-mak-2.0 až po vyše 1 200 000 textových jednotiek v r-mak-3.0, ktorý bude zrejme poslednou verziou ručne morfologicky anotovaného podkorpusu SNK vzhľadom na dosiahnutie štandardného rozsahu ručne anotovaných trénovacích a testovacích dát.

Osobitnou zložkou lingvistickej anotácie je syntaktická anotácia, ktorá posúva kvalitatívne parametre korpusu opäť vyššie – nie je takou štandardnou súčasťou ako lematizácia či morfologická anotácia. V SNK sa syntakticky anotuje podľa princípov funkčno-generatívneho opisu jazyka uplatneného pri tvorbe Pražského závislostného korpusu (http://ufal.mff.cuni.cz/pdt.html). Slovenský závislostný korpus (ďalej SZK) obsahuje texty, ktoré prešli ručnou morfologickou anotáciou, a každý text je anotovaný na prvej syntaktickej (analytickej) rovine dvoma rôznymi anotátormi (Šimková – Garabík, 2006; Gajdošová, 2006). V SZK sa momentálne nachádza vyše 50 tisíc dvakrát zanotovaných viet, presnejšie vetných celkov, ktoré po skorigovaní a sprístupnení verejnosti budú predstavovať ďalší dôležitý výskumný aj učebný materiál. Možnosť anotácie druhej syntaktickej – tektogramatickej roviny, ktorá zahŕňa valenčnú syntax, aktuálne členenie výpovede, konektorické vlastnosti a pod. (Anotace Pražského závislostního korpusu na tektogramatické rovině: pokyny pro anotátory, 2005) je v rámci slovenskej syntaktickej tradície zatiaľ v podstate nerealizovateľná.

Pracovisko SNK začalo s prípravou vlastného projektu Slovenskej terminologickej databázy (pracovná verzia s vyše 3 500 terminologickými záznamami je dostupná na adrese https://data.juls.savba.sk/std) zameranej na oblasť práva, ekonomiky a vybraných technických vied, ktorú dlhodobo žiada domáca prekladateľská i široká odborná verejnosť a v poslednom období ju potrebujú aj prekladatelia Európskej komisie, Európskeho parlamentu, ako aj iných európskych inštitúcií. V prípravnej fáze projektu bol vypracovaný návrh terminologického záznamu a návrh softvérovej štruktúry databázy s cvičnými údajmi (Levická, v tlači) a nadviazala sa spolupráca

s viacerými odborníkmi z oblasti ekonómie, práva, sociálnej práce a pod. SNK aj tu môže slúžiť ako zdroj odborných textov, z ktorých sa môžu automatizovane extrahovať termíny z jednotlivých oblastí.

Zo všetkých zložiek SNK je azda najznámejšou a najpriamejšie využívanou položka Lingvistické zdroje a slovníky. Verejnosti je v rámci tejto predovšetkým lexikografickej databázy on-line prístupných viacero slovníkov a pravopisných príručiek, ktoré slúžia nielen prakticky ako súčasné kodifikačné príručky, ale aj výskumne a teoreticky na poznanie vývinových tendencií slovenského jazyka (súčasťou databázy sú napr. pravopisné príručky z r. 1931 a 1940 a šesťzväzkový Slovník slovenského jazyka z r. 1959 – 1968). Všetky nové slovníky a nové vydania už existujúcich slovníkov vydávané v JÚĽŠ SAV po r. 1990 vznikli s počítačovou podporou (Benko, 2001) a slovenská lexikografia má aj vďaka tomu významné postavenie medzi lexikografickými školami vo svete.

3. Za prelom vo využívaní Slovenského národného korpusu a v modernej korpusovej lingvistike na Slovensku možno považovať projekt Morfosyntaktická analýza Slovenského národného korpusu, ktorý sa riešil v r. 2004 až 2006 za podpory grantovej agentúry VEGA pod vedením Miloslavy S o k o l o v e j v spoluriešiteľstve kolektívu SNK a pracovníkov filologických katedier Filozofickej fakulty Prešovskej univerzity v Prešove. Kým v predchádzajúcom období išlo o vybudovanie SNK so všetkými jeho zložkami vrátane základných lingvistických opisov a predovšetkým o využívanie na lexikografické ciele či o sporadické individuálne sondy do korpusu, v tomto projekte už išlo o systematickú kolektívnu prácu na získavaní a lingvistickom spracúvaní materiálu veľkého rozsahu, ako ho priebežne pripravoval a poskytoval Slovenský národný korpus, a bola to práca istým spôsobom priekopnícka a veľmi dynamická. Priekopnícka v zmysle vytvárania širšej koncepcie a metodiky lingvistického výskumu na materiáli dovtedy neexistujúcej elektronickej databázy slovenských textov, z ktorej je možné získať štatisticky presné a opakovane verifikovateľné výsledky, ako aj vytvárania koncepcie nového morfosyntaktického opisu súčasnej slovenčiny s využitím tých materiálových, metodologických i personálnych možností, ktoré boli k dispozícii a dali sa skĺbiť do zmysluplného celku. Dynamickosť spočívala okrem iného v prispôsobovaní sa budovaniu a stratégii sprístupňovania SNK, ktorý počas riešenia grantu vzrástol takmer dvojnásobne (vyvážený korpus, s ktorým autori pracovali najviac, takmer päťnásobne) a každý rok (spočiatku aj dvakrát ročne) bola zverejnená nová verzia doplnená nielen o nové texty, ale v niektorých prípadoch radikálne inovovaná aj z hľadiska lingvistickej (predovšetkým morfologickej, ale aj štýlovo-žánrovej) anotácie.

Výsledky výskumu sú zhrnuté v zborníku *Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli* (2006) obsahujúcom štúdie z korpusového výskumu najmä verbálnych lexém a tvarov, ktorým sa časť riešiteľského kolektívu venovala už pred začatím grantu, ale bez dostatočnej materiálovej bázy a aj v rámci metodologicky iného prístupu: valencia slovies a osobitne statických predikátov,

kondicionál prézenta a préterita, aktívne particípium préterita, transgresív, prospektívna konštrukcia *ísť* + infinitív. Rovnako aj ďalší autori nadviazali na svoje dlhodobejšie výskumné zámery a pripravili štúdie z oblasti výskumu anglických proprií, prevzatých slov, inštrumentálu plurálu maskulín a neohybných slovných druhov. Pokryli sa tak vybrané javy dvoch základných slovných druhov (pri slovesách v oveľa väčšom rozsahu ako pri podstatných menách) a poukázalo sa na problematickosť anotovania (formalizácie) a následne aj výskumu a opisu neohybných slovných druhov na základe materiálu z korpusu.

Predložené výsledky sú pri väčšine opisov slovesných javov vedené jasnou a jednotnou koncepciou vypracovanou vedúcou grantu a hlavnou zostavovateľkou zborníka M. Sokolovou (2006) a inšpirované korpusovou gramatikou anglického jazyka (Longman Grammar of Spoken and Written English). Samostatnými časťami niektorých príspevkov sú komunikačno-pragmatické aspekty používania daného tvaru. Takto spracované štúdie predstavujú dobré východisko na tvorbu syntetickej gramatiky slovenčiny na korpusovom základe. Ostatné štúdie sú založené viac na individuálnych prístupoch, s predchádzajúcimi ich zjednocuje najmä uvádzanie presných, pomerne jednotne spracovaných údajov o frekvencii dokladov: frekvencia morfosyntaktických prostriedkov v korpuse celkovo podľa štýlov (štýlová disperzia) a pomerná frekvencia v pomere k podielu daných textov vo vyváženom korpuse, ktorá má vyššiu výpovednú hodnotu než celková frekvencia. Uvádzanie a vysvetľovanie rozsiahlych kvantitatívnych zistení odlišuje tieto štúdie od všetkých doterajších lingvistických výskumov na Slovensku založených na vlastnej jazykovej introspekcii a ručnej excerpcii. Vzhľadom na nevyhnutnosť získavania a štatistického vyhodnocovania rozsiahlych údajov bolo potrebné aj zapojenie nelingvistických spolupracovníkov – informatikov, čím celá práca nadobudla výrazný interdisciplinárny rozmer.

Prínos projektu je zrejmý nielen v tom, že sa otvorila nová fáza korpusovolingvistického výskumu na Slovensku, ale i v tom, že sa pri riešení konkrétnych otázok ukázali možnosti aj obmedzenia práce s korpusom a poukázalo sa na potrebu skvalitňovania korpusu z hľadiska veľkosti a pestrosti (pre okrajové jazykové javy), z hľadiska dobrého výberu textov do vyváženého korpusu (pre frekventované jazykové javy) a doplnenia korpusu písaných textov o korpus hovorených komunikátov (na poznanie skutočne reálneho fungovania niektorých jazykových prostriedkov príznačných práve pre spontánnu, nepísomnú komunikáciu – Slovenský hovorený korpus sa už v súčasnosti buduje a do konca roka 2008 sa plánuje sprístupnenie jeho prvej verzie). Zároveň sa potvrdil prínos stratégie tvorby a sprístupňovania SNK v súvislosti s rôznymi podkorpusmi – okrem korpusu so všetkými textami (hlavný, najväčší korpus) je k dispozícii aj korpus menší, štýlovo vyvážený a s poslednými verziami aj samostatné podkorpusy podľa jednotlivých štýlov. Každý z nich sa osvedčil v tomto projekte pri inom type výskumu.

Z hľadiska vlastného lingvistického výskumu prinášajú prezentované výsledky nielen návrhy na modifikáciu a spresňovanie definícií v doterajších gramatických opi-

soch či pravopisných pravidlách, ale dopĺňajú aj pozorovania, ktoré sa predtým nerobili, resp. ktoré sa pred existenciou korpusu a nástrojov na získavanie a selekciu potrebných údajov ani robiť nedali. Za všetky príklad zo štúdie Jolany N i ž n í k o v e j (2006b), ktorá oproti predchádzajúcemu jednému významu a jednej valenčnej štruktúre slovesa *jestvovať* vymedzila na základe analýzy materiálu z korpusu 11 významov tohto slovesa s valenčnými štruktúrami. Dôležité sú aj zistenia o lexikalizácii niektorých tvarov v takom rozsahu, že by sa mali lexikograficky spracovať ako samostatné heslá. Praktické využitie pri výučbe slovenčiny (najmä ako cudzieho jazyka) môžu mať zistenia a závery o reálnom slovoslede v rámci niektorých (zložených) slovesných tvarov a možných rozptyloch medzi ich časťami (Kášová – Rabatinová, 2006; Kášová – Rabatinová – Dziak, 2006), ako aj o prípadoch systémovo neprotirečivého, no zatiaľ v takomto použití nekodifikovaného tvaru prípony *-mi* vs. *-ami* v inštrumentáli plurálu maskulín (Karčová – Majchráková, 2006), ak sa tieto výsledky teoreticky akceptujú a zahrnú do gramatických výkladov a pravopisných pravidiel.

Výsledkom precízne spracovaného a lingvisticky opísaného materiálu v štúdiách zhrnutých v zborníku Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli je aj poskytnutie vhodného východiska pre ďalšie korpusovolingvistické výskumy na Slovensku a pre syntetický opis gramatickej roviny súčasnej slovenčiny.

4. Počas piatich rokov budovania SNK sa podarilo vytvoriť národný korpus slovenského jazyka v dostatočne širokom rozsahu a zábere, vybavený výkonnými vyhľadávacími nástrojmi, nástrojmi na indexáciu a kategorizáciu textových dokumentov, nástrojmi na štatistickú analýzu a identifikáciu vývojových trendov v jazyku (rôzne štatistické údaje sú priamo dostupné na stránke korpusu). Využívanie SNK (každoročne vyše 200 registrovaných používateľov, neregistrovaní používatelia vstupujú do poskytovaných zdrojov v priemere vyše tridsaťtisíckrát denne) sa realizuje vo všetkých oblastiach, pre ktoré sa jazykové korpusy budujú:

a) Na Slovensku sa začala systematicky rozvíjať nová vedná disciplína – ***korpusová lingvistika*** ako odbor počítačovej lingvistiky, ktorej predmetom je skúmanie jazykových javov v prirodzených kontextoch vo veľkom množstve reálnych textov; na základe analýzy korpusových textov sa overujú doterajšie lingvistické teórie a môžu vzniknúť nové hypotézy a teórie, čoho dôkazom je aj grant Morfosyntaktická analýza Slovenského národného korpusu a zborník Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli (2006), 2. zväzok Valenčného slovníka slovenských slovies (na korpusovom základe) autorky J. N i ž n í k o v e j (2006a), monografie M. I v a n o v e j (2006) a M. S o k o l o v e j (2007), Slovník koreňových morfém slovenčiny (2006), ako aj viacero jazykovedne zameraných diplomových a dizertačných prác. Najvýznamnejšou aplikačnou zložkou v tejto oblasti je lexikografické využitie: mnoho korpusov sa budovalo a buduje na podporu tvorby slovníkov a lexikografi sú azda stále najčastejšími používateľmi korpusov – na báze materiálu Slovenského národného korpusu sa koncipuje nový 8-zväzkový Slovník súčasného slovenského jazyka (prvý zväzok A – G vyšiel v r. 2006) a aktualizujú sa jed-

notlivé vydania doterajších lexikografických i pravopisných príručiek, pripravuje sa frekvenčný, retrográdny a kolokačný slovník.

b) Niektoré výsledky zo spracovania korpusu, ako sú zoznamy slov, spoločné výskyty slov (kolokácie), frekvencia grafém, slabík, slov a spojení atď., sa používajú aj v nelingvistických aplikáciách – v neurológii, logopédii, psychológii, didaktike a pod., no predovšetkým v oblasti ***počítačového spracovania prirodzeného jazyka***, kde sa tvoria napr. systémy na spracovanie textov (automatická kontrola pravopisu, gramatiky či štylistiky, strojový preklad textov), systémy na rozpoznávanie reči a pod. – na materiáli Slovenského národného korpusu vzniklo a vzniká viacero diplomových a doktorandských prác na vysokých školách a výskumných pracoviskách technického zamerania.

c) Korpus môže byť a je dobrým zdrojom fráz a viet potrebných ***pri výučbe cudzieho, ale aj materinského jazyka***, využívajú ho učitelia doma i lektori slovenského jazyka v zahraničí na prípravu rôznych cvičení (porov. Ivoríková, 2006; Šimková, 2007); osobitne vítanou pomôckou v tejto oblasti sú paralelné korpusy, z ktorých sú sprístupnené zatiaľ tri (Parallel Corpus of Computer Terms, Francúzsko-slovenský paralelný korpus, Rusko-slovenský paralelný korpus) a postupne sa pripravujú ďalšie (slovensko-český, slovensko-anglický paralelný korpus).

Prehľad krátkej histórie matematickej lingvistiky na Slovensku, budovania Slovenského národného korpusu a rozvoja korpusovej lingvistiky na Slovensku v posledných piatich rokoch možno zhrnúť v troch časových bodoch: keď sa koncom 19. storočia realizovali prvé kvantitatívne analýzy svetových jazykov, spisovná slovenčina mala za sebou prvé desaťročia neľahkej existencie a nebolo isté, či a dokedy prežije; keď sa v 60. rokoch 20. storočia začínali vo svete budovať prvé jazykové korpusy a algebrická lingvistika prežívala druhé desaťročie plodného rozvoja, na Slovensku sa iba objavujú metódy matematickej lingvistiky a vydáva sa prvý frekvenčný slovník; keď v 90. rokoch 20. storočia nastupuje vo svete boom korpusov a korpusovej lingvistiky, v JÚĽŠ SAV sa pomaly začína s prvým malým interným korpusom textov slovenského jazyka bez lematizácie a lingvistickej anotácie. Na začiatku 21. storočia však už môžeme konštatovať, že zaostávanie v tejto oblasti na Slovensku bolo takmer eliminované. SNK sa dobre etabloval ako rýchlo sa rozvíjajúce pracovisko JÚĽŠ SAV a stal sa známym medzi korpusovými pracoviskami Európy, ako aj medzi domácimi a zahraničnými používateľmi korpusov z rôznych odborov. Medzinárodne porovnateľné výsledky dosahujú aj slovenskí odborníci na rozpoznávanie a syntézu reči, ktorí dlhodobo pracujú v tomto samostatnom odbore počítačového spracovania prirodzeného jazyka predovšetkým v Ústave informatiky Slovenskej akadémie vied, ale aj v komerčnej sfére (porov. štúdie v zborníkoch z medzinárodných konferencií Slovko 2001; 2005; 2007). Pozitívny prínos SNK v oblasti rozvoja jazykových a informačných technológií formulujeme predovšetkým z lingvistického hľadiska; zlepšenie ratingu slovenskej korpusovej lingvistiky a počítačového spracovania slovenčiny je však otvorené a perspektívne bude zrejme vyžadovať väčšie zaangažovanie inžinierskych a informatických zložiek.

Bibliografia

ALTMANN, Gabriel: Kvantitatívne metódy v literárnej vede. In: Slovenská literatúra 1963, roč. 10, s. 432 – 441.

ALTMANN, Gabriel: O statističeskich metodach issledovanija malajskoj literatury. In: Narody Azii i Afriki 1969, s. 137 – 141.

ALTMANN, Gabriel: Status und Ziele der quantitativen Sprachwissenschaft. In: Jäger, S. (Hrsg.), Linguistik und Statistik. Braunschweig: Vieweg 1972, s. 1 – 9.

Anotace Pražského závislostního korpusu na tektogramatické rovině: pokyny pro anotátory. Praha: Matematicko-fyzikální fakulta Univerzity Karlovy 2005. 1202 s. Dostupný aj z WWW: http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch06.html

BENKO, Vladimír: Počítačová podpora lexikografických projektov – retrospektívny pohľad. In: Slovenčina a čeština v počítačovom spracovaní. Zborník referátov zo seminára Slovko 2001. Ed. A. Jarošová. Bratislava: Veda 2001, s. 181 – 194.

Computer Treatment of Slavic and East European Languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2005. Ed. R. Garabík. Bratislava: Veda 2005. 246 s. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/slovko2005/

Computer Treatment of Slavic and East European Languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2007. Ed. J. Levická – R. Garabík. Bratislava: Tribun 2007. 318 s. Dostupný aj z WWW: http://korpus.juls.savba.sk/~slovko/2007/Slovko_2007_proceedings.pdf

ČERMÁK, František: Jazykový korpus: prostředek a zdroj poznání. In: Studie z korpusové lingvistiky. Acta Universitatis Carolinae Philologica 3 – 4, 1997. Praha: Nakladatelství Karolinum 2000, s. 15 – 37.

ČERMÁK, František: Český národní korpus: stav v roce 2001. In: Slovenčina a čeština v počítačovom spracovaní. Zborník referátov zo seminára Slovko 2001. Ed. A. Jarošová. Bratislava: Veda 2001, s. 121 – 135.

ČERMÁK, František: Korpusová lingvistika dnešní doby. In: Korpusová lingvistika: Stav a modelové přístupy. Studie z korpusové lingvistiky. Sv. 1. Ed. F. Čermák – R. Blatná. Praha: Nakladatelství Lidové noviny – Ústav Českého národního korpusu 2006, s. 9 – 18.

ČERNÝ, Jiří: Dějiny lingvistiky. Olomouc: Votobia 1996. 512 s.

DOMIN, Pavol: Niektoré technické aspekty budovania a používania korpusov. In: Slovenčina na začiatku 21. storočia. Ed. M. Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, s. 159 – 163. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2003

FORRÓOVÁ, Martina – GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária: Návrh morfologického tagsetu SNK. In: Slovanské jazyky v počítačovom spracovaní. Slovko 2003. Materiály z medzinárodnej konferencie neboli publikované. Dostupný z WWW: http://korpus.juls.savba.sk/publications/block2/2005

FORRÓOVÁ, Martina – HORÁK, Alexander: Morfologická anotácia korpusu. In: Slovenčina na začiatku 21. storočia. Ed. M. Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, s. 174 – 186. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2003

GAJDOŠOVÁ, Katarína: Syntaktická anotácia vybraných textov Slovenského národného korpusu. XVI. kolokvium mladých jazykovedcov, Častá-Papiernička 8. – 10. 11. 2006. (v tlači) Dostupný z WWW: http://korpus.juls.savba.sk/publications/block2/2006

GARABÍK, Radovan: Štruktúra dát v Slovenskom národnom korpuse a ich vonkajšia anotácia. In: Slovenčina na začiatku 21. storočia. Ed. M. Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, s. 164 – 173. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2003

GARABÍK, Radovan: Corpus Construction Tools. In: Труды международной конференции MegaLing'2005. Прикладная лингвистика в поиске новых путей. Zborník z medzinárodnej vedeckej konferencie MegaLing'05. Meganom, Krym, Ukrajina 27. júna – 2. júla 2005. Red. В. П. Захаров – С. С. Дикарева. С.-Петербург: Издательство «Осипов» 2005, s. 26 – 32. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2005

GARABÍK, Radovan: Levenshtein Edit Operations as a Base for a Morphology Analyzer. In: Computer Treatment of Slavic and East European Languages. In: Zborník z medzinárodnej vedeckej konferencie Slovko 2005. Red. R. Garabík. Bratislava: Veda 2005, s. 50 – 58. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2006

GARABÍK, Radovan: Slovak morphology analyzer based on Levenshtein edit operations. In: 1st Workshop on Intelligent and Knowledge oriented Technologies. Proceedings of the WIKT'06 conference. Ed.: M. Laclavík – I. Budinská – L. Hluchý. Bratislava: Institut of Informatics Slovac Academy of Sciences 2007, pp. 2 – 5. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2006

GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. 2004. Interný materiál. Dostupný z WWW: http://korpus.juls.savba.sk/publications/block2/2004

GARABÍK, Radovan – GIANITSOVÁ-OLOŠTIAKOVÁ, Lucia: Manual Morphological Annotation of the Slovak Translation of Orwell's Novel 1984 – Methods and Findings. In: Computer Treatment of Slavic and East European Languages. In: Zborník z medzinárodnej vedeckej konferencie Slovko 2005. Ed. R. Garabík. Bratislava: Veda 2005, s. 59 – 66. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2005

HORECKÝ, Ján: Úvod do matematickej jazykovedy. Bratislava: Rektorát Univerzity Komenského 1969. 92 s.

HORECKÝ, Ján: Projekt bázy dát slovenského jazyka. In: Metódy výskumu a opisu lexiky slovanských jazykov. Materiály zo sympózia konaného v rámci 7. zasadnutia Lexikologicko-lexikografickej komisie pri Medzinárodnom komitéte slavistov (Nové Vozokany 24. – 26. apríla 1989). Red. V. Blanár et al. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV 1990, s. 251 – 253.

Insight into the Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2006. 208 s. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/insight/

IVANOVÁ, Martina: Valencia statických slovies. Prešov: Filozofická fakulta Prešovskej univerzity 2006. 273 s.

IVORÍKOVÁ, Helena: Využitie Slovenského národného korpusu (SNK) v školskej praxi (SNK ako pomôcka pri vyučovaní slovenského jazyka). Dostupný z WWW: http:// korpus.juls.savba.sk/publications/block2/2006.

KARČOVÁ, Agáta – MAJCHRÁKOVÁ, Daniela: Distribúcia koncoviek *-mi, -ami* v tvaroch maskulínnych substantív zakončených na konsonantické skupiny. In: Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Ed. M. Sokolová – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove 2006, s. 155 – 175. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2006

KARČOVÁ, Agáta – ŠIMKOVÁ, Mária: Морфологічна анотація текстів словацького національного корпусу. In: Лексикографічний бюлетень 13. Київ: Інститут української мови Національної академії наук України 2006, s. 71 – 76. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2006

KÁŠOVÁ, Martina – RABATINOVÁ, Slavomíra: Výskum kondicionálu prézenta. In: Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Ed. M. Sokolová – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove 2006, s. 40 – 60.

KÁŠOVÁ, Martina – RABATINOVÁ, Slavomíra – DZIAK, Vladimír: Výskum kondicionálu préterita v slovenčine. In: Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Ed. M. Sokolová – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove 2006, s. 61 – 76.

KILGARRIFF, Adam – RYCHLY, Pavel – SMRZ, Pavel – TUGWELL, David: The Sketch Engine. In: EURALEX 2004. Dostupný z WWW: http://trac.sketchengine.co.uk/attachment/wiki/SkE/DocsIndex/sketch-engine-elx04.pdf?format=raw

KRÁLÍK, Jan: Aktualizace rozvržení zdrojů Českého národního korpusu s ohledem na revizi vyváženosti jeho struktury. In: Slovo a slovesnost, 2004, roč. 65, č. 2, s. 133 – 141.

Krátky slovník slovenského jazyka. 4., doplnené a upravené vyd. Pripravili a zredigovali J. Kačala – M. Pisárčiková – M. Považaj. Bratislava: Veda 2003. 985 s.

LEVICKÁ, Jana: Slovenská terminologická databáza. In: Kultúra slova, 2008. (v tlači)

MISTRÍK, Jozef: Frekvencia slov v slovenčine. Bratislava: Vydavateľstvo SAV 1969. 726 s.

MISTRÍK, Jozef: Retrográdny slovník slovenčiny. Bratislava: Univerzita Komenského 1976. 736 s.

MISTRÍK, Jozef: Frekvencia tvarov a konštrukcií v slovenčine. Bratislava: Veda 1985. 320 s.

NIŽNÍKOVÁ, Jolana: Valenčný slovník slovenských slovies. 2. zväzok (na korpusovom základe). Prešov: Filozofická fakulta Prešovskej univerzity 2006(a). 166 s.

NIŽNÍKOVÁ, Jolana: Výskum valencie slovies. In: Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Ed. M. Sokolová – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove 2006(b), s. 114 – 118.

OLIVA, Karel – DOLEŽALOVÁ, Drahomíra „Johanka“: O korpusu jako o zdroji jazykových dat. In: Korpus jako zdroj dat o češtině. Ed. P. Karlík. Brno: Masarykova univerzita v Brně 2004, s. 7 – 10.

Slovenčina a čeština v počítačovom spracovaní. Zborník referátov zo seminára Slovko 2001. Ed. A. Jarošová. Bratislava: Veda 2001. 196 s. Dostupný aj z WWW: http://korpus.juls.savba.sk/files/zbornik/sacvps.pdf

Slovník koreňových morfém slovenčiny. Red. M. Sokolová – M. Ološtiak – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove 2005. 584 s.

Slovník súčasného slovenského jazyka. A – G. Red. K. Buzássyová – A. Jarošová. Bratislava: Veda 2006. 1134 s.

SOKOLOVÁ, Miloslava: Options for the Generation of a Corpus-Based Slovak Morphology (as Part of Corpus Morphosyntax). In: Insight into the Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2006, s. 179 – 205.

SOKOLOVÁ, Miloslava: Nový deklinačný systém slovenských substantív. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove 2007. 338 s.

Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Ed. M. Sokolová – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove 2006. 196 s.

ŠIMKOVÁ, Mária: Počítačové spracovanie prirodzeného jazyka a Slovenský národný korpus. In: Počítačová podpora prekladu. Zborník prednášok. Ed. M. Smolík – J. Šoltys – F. Tomášik. Bratislava: Slovenská spoločnosť prekladateľov odbornej literatúry 2003, s. 15 – 19. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2003

ŠIMKOVÁ, Mária: Slovenský národný korpus – východiská a plány. In: Slovenčina na začiatku 21. storočia. Ed. M. Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004(a), s. 150 – 158. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2004

ŠIMKOVÁ, Mária: Možnosti využitia Slovenského národného korpusu na štúdium slovenského jazyka. In: Studia Academica Slovaca 33. Prednášky z XL. letnej školy slovenského jazyka a kultúry. Ed.: J. Mlacek – M. Vojtech. Bratislava: Filozofická fakulta Univerzity Komenského 2004(b), s. 204 – 218. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2005

ŠIMKOVÁ, Mária: Репрезентативность корпуса как лингвистическая проблема. In: Труды международной конференции MegaLing'2005. Прикладная лингвистика в поиске новых путей. Zborník z medzinárodnej vedeckej konferencie MegaLing'05. Meganom, Krym, Ukrajina 27. júna – 2. júla 2005. Red. V. P. Zacharov – S. S. Dikareva. St. Petersburg: Izdateľstvo Osipov 2005, s. 130 – 139. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2005

ŠIMKOVÁ, Mária – GARABÍK, Radovan: Синтаксическая разметка в Словацком национальном корпусе. In: Труды международной конференции Корпусная лингвистика – 2006. Sankt-Petersburg: St. Petersburg University Press 2006, s. 389 – 394. Dostupný aj z WWW: http://korpus.juls.savba.sk/publications/block1/2006

ŠMOTLÁK, Martin – HLADÍK, Rudolf: Conversion of Microsoft Word and OpenOffice into XML-like documents. Text nebol publikovaný. Dostupný z WWW: http://korpus.juls.savba.sk/publications/block2/2004