

DESAŤ ROKOV NÁRODNÉHO KORPUSU

Desať rokov nie je nijako veľa ani v živote človeka a už tobôž nie vo vedeckom živote. Čo však v štandardných podmienkach vyzerá ako normálny jav, ktorému netreba venovať osobitnejšiu pozornosť, v situácii permanentného ohrozovania základnej existencie, ktorá je príznačná pre slovenské vedecko-výskumné pracoviská osobitne v oblasti spoločenských a humanitných vied, si predsa len zaslúži istú pozornosť. Doterajší príbeh Slovenského národného korpusu môžeme zároveň označiť ako príklad reagovania slovenskej vedy a výskumu na medzinárodné trendy, na potreby verejnosti (širokého spektra domácich i zahraničných bádateľov a iných záujemcov) a príklad úsilia o vyplnenie bieleho miesta vo vedeckom výskume na Slovensku.

Osemdesiate roky a začiatok deväťdesiatych rokov 20. storočia boli (nielen) v slovenskej spoločnosti poznamenané zvýšeným tempom rozvoja, ktorého odraz v jazyku nevyhnutne zachytávala aj jazykoveda. Išlo jednak o zmenu vedeckej paradigmy v rámci komunikačno-pragmatického obratu a masívnejšieho nástupu interdisciplinárnych odborov (sociolingvistika, etnolingvistika, psycholingvistika, textová lingvistika, filozofia jazyka, ale aj matematická, kvantitatívna a počítačová lingvistika), jednak o opisy slovnej zásoby, ktorá najcitlivejšie reaguje na spoločenské pohyby. V jazykovednom ústave L. Štúra SAV sa v tom čase realizovali viaceré projekty dosahujúce svojím rozsahom i vedeckou erudíciou medzinárodné uznanie a iniciovali sa nové výskumné zábery, ktoré dodnes poznamenávajú vedecko-výskumné smerovanie pracoviska. Z vtedy ukončených projektov máme na mysli *Krátky slovník slovenského jazyka* (1. vyd. 1987, 2. vyd. 1989), zachytávajúci slovnú zásobu „*obdobia budovania socialistickej spoločnosti (zhruba obdobie posledných 40 rokov)*“, ktorého autori v 3. vydaní (1997) prepracovali asi tretinu pôvodného rukopisu: doplnili nové slová a významy, niektoré heslá vyradili, viaceré významy zmenili zodpovedajúco novej spoločensko-politickej situácii. Dve vydania toho istého opisu slovnej zásoby na priesečníku diania tak poskytujú vynikajúci materiál na porovnanie jazyka pred r. 1989 a krátko po ňom (porov. Šimková, M.: Pohyby v slovenskej lexike v 90. rokoch. In: *Internacionalizácia v súčasných slovanských jazykoch: za a proti*. Red. J. Bosák. Bratislava: Veda 1999, s. 116 – 136). Podobne významným projektom bola monografia *Dynamika slovnej zásoby súčasnej slovenčiny* (Horecký, J. – Buzássyová, K. – Bosák, J. a kol., Bratislava: Veda 1989), v ktorej autori mapovali slovnú zásobu a nové tendencie v slovenčine v 70. a 80. rokoch 20. storočia. Obidve diela, ako aj ďalšie špecializované slovníky

a gramatické práce sa koncipovali na materiálnej báze kartotéky/kartoték. V nich sa nachádza vyše 5 miliónov excerpčných lístkov s výpiskami krátkych kontextov istého výberu slovenských slov uložených podľa abecedy v neprenosných škatuliach v budove JÚLŠ SAV.

Začiatkom 90. rokov sa excerpčia musela pre nedostatok finančných prostriedkov zastaviť. Na našich pracovných stoloch sa však začali objavovať počítače – sprvu najmä ako sponzorské dary alebo súkromne zakúpené niekoľkými nadšencami z vlastných prostriedkov. Súčasne k nám v širšom rozsahu prúdili informácie o nových možnostiach jazykovednej, najmä lexikografickej práce na báze elektronických databáz, ktoré už od 60. rokov 20. storočia rozvíjali americkí a anglickí počítačoví experti, lingvisti a lexikografi a postupne sa rozširovali aj na ďalšie jazyky vrátane krajín strednej a východnej Európy. Záujem odborníkov v oblasti počítačového spracovania prirodzeného jazyka o slovanské jazyky podnietil potrebu formálnych, matematických opisov flektívnych jazykov, na čo sa priebežne reagovalo aj na Slovensku: známe sú v tomto smere práce a aktivity J. Horeckého, neskôr monografia E. Páleša, ako aj čiastkové výsledky pracovníkov bývalého Ústavu technickej kybernetiky SAV E. Kostolanského a V. Benka.

Začatie projektu tvorby nového veľkého výkladového slovníka súčasného slovenského jazyka (známeho SSSJ, z ktorého sú na trhu zatiaľ zväzky A – G a H – L) v r. 1993 zvýšilo tlak na vytvorenie modernej a rozsiahlej materiálnej bázy. Kvantitatívne aj kvalitatívne obmedzená kartotéka, ktorej dopĺňanie bolo aj tak zastavené, nemohla takéto požiadavky splniť. V JÚLŠ SAV sa preto pristúpilo k budovaniu interného korpusu textov slovenského jazyka ako elektronickej materiálnej databázy. A opäť to bolo len pár aktívnych nadšencov, ktorí popri inej svojej práci získavali východiskové texty v elektronickej podobe (niekoľko ročníkov novín a časopisov, niekoľko kníh), k nim sa pridali elektronicke verzie slovníkov z produkcie JÚLŠ SAV, počítačové nástroje na spracovanie textov a vyhľadávanie v nich sa vyberali z voľne dostupných programov (*shareware*), základné štruktúrovanie a úplne jednoduchá bibliografická anotácia sa robili prevažne ručne. Takáto situácia nebola dlhodobo udržateľná a dosahované výsledky zďaleka nezodpovedali potrebám domáceho výskumu ani vývoju vo svete. Po asi šiestich rokoch mal tento prototyp slovenského korpusu iba cca 25 mil. slov, lingvistická anotácia v ňom nebola žiadna, pretože na to neboli vyčlenené nijaké pracovné úväzky, šerverové programy prakticky neumožňovali ďalej zväčšovať rozsah dát ... a začínali sa rozbiehať



VÝSLEDKY SVOJEJ PRÁCE PREZENTOVALO ODDELENIE SNK JÚLŠ SAV NA MEDZINÁRODNEJ KONFERENCII, 7. – 8. JÚNA 2012 V BRATISLAVE. POHĽAD NA PRACOVNÉ PREDSEDNÍCTVO KONFERENCIE – ZĽAVA: JAN HAJIČ, LUBOMÍR FAĽTAN, PAVOL ŽIGO, MARTA CIMBÁKOVÁ, TAMÁS VÁRADI, GEORG REHM. Foto Juraj Broš

prístupové konania do Európskej únie, do ktorej Slovensko chcelo vstúpiť ako vyspelá krajina rovnocenná s ostatnými vo všetkých oblastiach a ukazovateľoch. Súhrn vonkajších a vnútorných faktorov, ktoré stáli pri zrode osobitného pracoviska Slovenského národného korpusu v JÚLŠ SAV, doplnili ustanovenia zákona o štátnom jazyku zdôrazňujúce o. i. úlohu štátu pri starostlivosti o poznanie a rozvoj slovenčiny ako plne funkčného samostatného jazyka v rodine všetkých európskych i svetových jazykov.

Príprava a schvaľovací proces projektu *Výbudovanie Národného korpusu slovenského jazyka a elektronizácia jazykovedného výskumu v rokoch 2002 – 2006* napriek akútnym potrebám a všetkým splneným podmienkam trvali takmer dva roky. Vďaka Ministerstvu kultúry SR, Ministerstvu školstva SR a vedeniu SAV sa napokon 13. februára 2002 dospelo k prijatiu vládneho uznesenia č. 137, ktorým sa zabezpečilo financovanie zriadenia oddelenia Slovenského národného korpusu JÚLŠ SAV a jeho fungovania na 5 rokov. V podmienkach, keď akademické ústavy museli šetriť finančné prostriedky prepúšťaním pracovníkov, znižovaním spotreby elektrickej energie a radikálnym obmedzením všetkých ďalších výdavkov, boli začiatky nového pracoviska naozaj grandiózne: rekonštrukcia strechy, výstavba a kompletne zariaďovanie nových podkrovných priestorov v budove JÚLŠ SAV, konkurzy a prijímanie nových pracovníkov, plány na vybudovanie špecifických databáz (lexikografická, terminologická) a všetkých druhov vtedy známych korpusov: písaný, paralelný, hovorený, nárečový, historický... Sedemdesiaty kolektív prevažne mladých ľudí, ktorí sa dovtedy s počítačovou či korpusovou lingvistikou stretli prevažne iba

okrajovo, stál pred mimoriadne náročnou úlohou! V príprave na jej plnenie boli našťastie k dispozícii ústretoví kolegovia z už existujúcich pracovísk podobného typu v okolitých krajinách, najmä v Českej republike – na pôde SNK JÚLŠ SAV prebiehali odborné prednášky a semináre pod ich vedením, časté boli aj pracovné návštevy v Brne (Fakulta informatiky MU) či v Prahe, kde sú tri veľké pracoviská počítačovej, formálnej a korpusovej lingvistiky európskeho, ba aj svetového významu. Získavanie textov sa najprv muselo právne zabezpečiť, na ich spracúvanie sa implementovali zahraničné nástroje a programové vybavenia. V priebehu troch rokov bolo sprístupnených 5 verzií Slovenského národného korpusu (<http://korpus.juls.savba.sk>), ktorého celkový rozsah dosiahol 300 miliónov textových jednotiek, prvé paralelné korpusy (slovensko-ruský a slovensko-francúzsky), v elektronickej podobe sa dali k dispozícii širokej verejnosti základné lingvistické, najmä lexikografické diela (<http://slovníky.korpus.sk/>), boli vytvorené viaceré vlastné počítačové nástroje vrátane konverzných programov, lematizátora a morfológického analyzátora, pracovisko SNK JÚLŠ SAV sa začalo zapájať do medzinárodných spoluprác, zorganizovalo medzinárodnú konferenciu a za svoju činnosť a výsledky bolo v r. 2005 ocenené Cenou SAV za budovanie infraštruktúry pre vedu.

Veľkou výhodou pri budovaní a skvalitňovaní databáz SNK, ktorú nám v dobrom závideli aj zahraniční kolegovia, boli „odberatelia“. Išlo predovšetkým o pracovníkov lexikografického kolektívu nového výkladového slovníka v JÚLŠ SAV a pracovníkov vtedajšej Katedry slovenského jazyka a literatúry Filozofickej fakulty v Prešove, ale aj iných, ktorí jednak poskytovali spätnú väzbu formulovaním konkrétnych skúseností z využívania korpusového materiálu a upozorňovaním na problémové miesta, jednak sa dožadovali ďalších a ďalších elektronických zdrojov potrebných na realizáciu ich výskumných a vzdelávacích cieľov. Odpoveďou na tieto požiadavky a v súlade s celosvetovým rozvojom korpusovej lingvistiky a počítačového spracovania prirodzených jazykov vrátane rozvoja jazykových a informačných technológií bolo vypracovanie projektu *Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku (druhá etapa)*, ktorý bol schválený a finančne podporený opäť ministerstvami kultúry a školstva i Akadémiou. Od r. 2007 sa tak mohlo bez prerušenia pokračovať v rozširovaní a skvalitňovaní už vytvorených databáz a začať s tvorbou a sprístupňovaním nových elektronických zdrojov.

Skvalitňovanie sa týkalo najmä lingvistickej anotácie (ručnej morfológickej anotácie, ku ktorej pribudla morfológická databáza v rozsahu približne 75 000 slov s vyše 2 mil. tvarov), konverzných nástrojov a nástrojov na automatické zarovnávanie textov, ako i webového rozhrania na prácu s paralelnými korpusmi. Rozširovanie primárneho korpusu písaných textov sa aj od r. 2007 realizovalo podobne ako v predchádzajúcom období, t. j. pri každom rele-

vantnom náraste dát a/alebo skvalitnení anotácie bola sprístupnená nová verzia korpusu ako rozšírenie predchádzajúcej verzie: začiatkom r. 2007 to bol korpus prim-3.0 v rozsahu 350 miliónov tokenov (textových jednotiek), začiatkom r. 2009 prim-4.0 v rozsahu 526 mil. tokenov, začiatkom r. 2011 prim-5.0 v rozsahu 719 mil. tokenov. S každou verziou boli sprístupnené samostatné podkorpusy publicistických textov, odborných textov, umeleckých textov a pôvodných slovenských umeleckých textov. Pre záujemcov o teoretické poznanie slovenčiny bolo veľmi dôležité aj rozšírenie elektronických lexikografických príručiek a lingvistických zdrojov, medzi ktoré pribudli všetky ročníky časopisov vydávaných v JÚLŠ SAV (*Kultúra slova* a jej predchodca *Československý terminologický časopis*, *Slovenská reč*, *Jazykovedný časopis*), viaceré zborníky a monografie, prvé Pravidlá slovenského pravopisu (z r. 1931 a 1940) a najmä 5-väzkový slovník A. Bernoláka z r. 1825. Nové korpusové databázy sa od r. 2007 budovali nielen v rámci písaných textov – pribudol samostatný korpus textov z r. 1955 – 1989 (druhá verzia má 44 mil. tokenov), slovensko-český (cca 20 mil. tokenov) a slovensko-anglický paralelný korpus (cca 45 mil. tokenov), webový korpus (1 miliarda tokenov) –, ale po osobitnej teoretickej príprave sa pristúpilo aj k tvorbe Slovenského hovoreného korpusu (<http://korpus.sk/shk.html>), v ktorom sa zachytáva štandardná hovorená slovenčina z celého územia Slovenska. Prvé verzie hovoreného korpusu boli sprístupnené v r. 2008 a 2010, tretia verzia v r. 2011, aktuálna štvrtá verzia z r. 2012 je v rozsahu 2 611 000 slov. V porovnaní s miliardovými korpusmi písaných textov je to na prvý pohľad veľmi malý počet jednotiek, ktorý však predstavuje v súčasnosti viac ako štandardný rozsah hovoreného korpusu. Náročnosť jeho tvorby a pomalšie pribúdanie dát je spôsobené podstatou hovoreného jazyka, ktorý sa musí najskôr zaznamenať (niekedy je náročnejšie získať niekoľko minútový rozhovor ako celý ročník časopisu), musí sa slovo po slove ručne prepisovať do textovej podoby, segmentovať, anotovať a viacstupňovo opravovať. Samostatným podprojek-

tom v rámci komplexu SNK je Slovenská terminologická databáza (<http://data.juls.savba.sk/std/>), v ktorej sa nachádza približne 5 000 terminologických záznamov z 20 oblastí. Členovia oddelenia SNK sa v druhej etape svojej činnosti (2007 – 2011) zapojili do štyroch významných medzinárodných projektov (MONDILEX, Slovak Online, EuroMatrix+, CESAR), spolupracovali na štyroch domácich projektoch (<http://korpus.sk/projects.html>), zorganizovali tri medzinárodné a dve domáce konferencie, z ktorých vydali zborníky. Databázy Slovenského národného korpusu a súvisiacich projektov poskytujú širokej odbornej i laickej verejnosti bohatý textový a jazykový materiál, na ktorom sa realizujú rôzne výskumné práce v spoločenských, informatických a iných technických a prírodných vedách a ich výsledky sa publikujú v podobe čiastkových štúdií i syntetických monografií a slovníkov. Počet registrovaných používateľov korpusových a lingvistických zdrojov stúpol na približne 500 ročne (najväčší nárast sme zaregistrovali z prostredia učiteľov základných a stredných škôl), počet neregistrovaných dotazov sa pohybuje v priemere okolo 40 000 denne.

Výsledky svojej práce prezentovali členovia oddelenia SNK JÚLŠ SAV aj na ďalšej medzinárodnej konferencii, ktorú zorganizovali 7. – 8. júna 2012 v Bratislave pod názvom *Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu)* (<http://korpus.sk/snk10.html>). Toto vedecko-informačné podujatie konané pod záštitou predsedu SAV J. Pastoreka pripravili v rámci prezentačných dní partnerských krajín projektu CESAR (Maďarsko – koordinátor, Poľsko, Slovensko, Chorvátsko, Srbsko, Bulharsko), ktorého zámerom je sprístupniť existujúce jazykové zdroje a zapojiť sa do tvorby jednotného digitálneho trhu v Európe. Mapovanie stavu elektronických databáz a jazykových technológií sa realizovalo spoločne v rámci projektov CESAR a META-NET prostredníctvom série bielych kníh. Postavenie Slovenska v kontexte ostatných zapojených európskych krajín na základe údajov zhrnutých v publikácii Šimková, M. a kol.: *Slovenský jazyk v digitálnom veku*

POHĽAD DO KONFERENČNEJ SÁLY.



(Berlín : Springer Verlag 2012, 92 s.) hodnotili na konferencii najmä Georg Rehm z Nemeckého výskumného centra umelej existencie v Berlíne a Tamás Váradi z Maďarskej akadémie vied. Okrem Slovenského národného korpusu (Mária Šimková, Radovan Garabík) boli prezentované aj ďalšie všeobecné národné a špecializované korpusy: Chorvátsky národný korpus, Český národný korpus, Národný korpus záverečných prác slovenských vysokých škôl, Maďarský národný korpus. Možnosti a nástroje na tvorbu a používanie jazykových zdrojov predstavovali prednášajúci zo Slovenskej národnej knižnice v Martine, z Ústavu informatiky SAV, z Fakulty informatiky Masarykovej univerzity, z Technickej univerzity v Košiciach a z Bulharskej akadémie vied. Osobitný projekt na zber jazykových dát z oblasti humanitných vied prezentoval Jan Hajič z Matematicko-fyzikálnej fakulty KU v Prahe (*LINDAT-Clarin, jazyková infraštruktúra pre výskum*). Súčasťou konferencie bola posterová sekcia a samostatné prezentácie firiem pôsobiacich v oblasti jazykových a informačných technológií.

Keďže jednou z hlavných myšlienok národných prezentačných dní je spojenie všetkých zainteresovaných z komerčnej sféry, štátnych a vzdelávacích inštitúcií, vedecko-výskumných centier na tvorbe a zdokonaľovaní jazykových technológií, v rámci panelovej diskusie v prvý deň rokovania sa otvorene preberali potreby, možnosti a prekážky takejto spolupráce na Slovensku. Plnú podporu rozvoju počítačového spracovania slovenčiny vyslovili v rámci slávnostného otvorenia podpredseda SAV Ľubomír Falťan i generálna riaditeľka sekcie vedy a techniky MŠVVaŠ SR Marta Cimbáková, no na zlepšenie možností využitia výsledkov práce oddelenia SNK JÚLŠ SAV a ďalších vedecko-výskumných pracovísk podobného charakteru v komerčnej sfére budú potrebné o. i. viaceré zmeny v legislatíve. Slovenský národný korpus vznikol a celé desaťročie fungoval ako vedecko-výskumný, nekomerčný projekt, ktorého databázy tvoria texty poskytnuté na základe licenčných zmlúv bezplatne práve so záväzkom nekomerčného využitia. Textové zdroje však boli technicky prepracované, boli do nich vnesené dodatočné lingvistické informácie a vytvorili sa rôzne pomocné databázy na čo najlepšie výsledné spracovanie textov. Pre ďalší vývoj jazykových a informačných technológií pre slovenčinu sú zaujímavé práve tieto pomocné databázy a konečné výsledky spracovania textov (frekvenčné zoznamy písmen, slov, spojení a pod.). V tretej etape projektu Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku (2012 – 2016) sa riešitelia zatiaľ opäť zamerajú predovšetkým na potreby vedecko-výskumných a vzdelávacích projektov, na rozšírenie existujúcich zdrojov a tvorbu špecializovaných korpusov (nárečové, historické texty, paralelné slovensko-nemecké, slovensko-maďarské texty) a príručiek.

Mária Šimková | Foto Juraj Broš