Aktuálny zdroj jazykovej informácie

Ďalšia fáza budovania Národného korpusu slovenského jazyka

V roku 2002 bol v 9. čísle Správ SAV publikovaný príspevok A. Jarošovej Národný korpus slovenského jazyka a jeho dimenzie, v ktorom autorka informovala o jednom z dôležitých fenoménov súčasnej svetovej lingvistiky a rozvoja informačných technológií – o jazykových korpusoch. Upozornila aj na nevyhnutnosť nájsť osobitné zdroje na budovanie elektronického korpusu textov slovenského jazyka v súvise s potrebou adekvátnej materiálovej bázy pre koncipovanie nového výkladového slovníka súčasnej slovenčiny (jeho prvý zväzok A – G vyšiel vo VEDE roku 2006), pre aktualizáciu existujúcich lexikografických a pravopisných príručiek (Krátky slovník slovenského jazyka, Pravidlá slovenského pravopisu a pod.) i gramatických diel (morfologický a syntaktický opis slovenčiny) a, ako autorka zdôraznila v závere, nie je to len „vytvorenie objektívneho a autentického zdroja jazykovej informácie na tvorbu základných akademických diel a aktualizácia jestvujúcich jazykových príručiek", ale korpus jazyka je aj „komponentom informatizácie našej spoločnosti a súčasne dôležitým predpokladom prežitia jazyka počtom malého národa v budúcej informačnej spoločnosti". Budovanie Národného korpusu slovenského jazyka v Jazykovednom ústave Ľ. Štúra SAV a elektronizácia jazykovedného výskumu na Slovensku sa stali skutočnosťou na základe uznesenia vlády Slovenskej republiky č. 137 z 13. 2. 2002 podľa návrhu Slovenskej akadémie vied, Ministerstva školstva SR a Ministerstva kultúry SR, ktoré je ústredným orgánom štátnej správy v oblasti starostlivosti o štátny jazyk. Projekt bol v predkladanej podobe schválený do r. 2006, od 1. 1. 2007 pokračuje Budovanie Slovenského národného korpusu a elektronizácia jazykovedného výskumu na Slovensku (druhá etapa) na základe osobitnej zmluvy medzi MŠ SR, MK SR a SAV. Aké boli dosiahnuté výsledky a aké sú plány na ďalšie obdobie?

Veľkej časti nielen lingvistickej počítačovej verejnosti na Slovensku a záujemcom zo zahraničia je už dôverne známa www stránka JÚĽŠ SAV (http://www.juls.savba.sk), a to najmä vďaka ponuke lingvistických zdrojov a slovníkov. Bezplatne prístupné sú tu elektronické verzie aktuálnych vydaní základných kodifikačných príručiek (Krátky slovník slovenského jazyka, Pravidlá slovenského pravopisu), ale aj starší slovník väčšieho rozsahu (Slovník slovenského jazyka, 1959 – 1968), ktorý poskytuje obraz o vývine, dynamike súčasnej slovenčiny a obsahuje aj množstvo dodnes platných spisovných slov a tvarov. Medzi lexikografické príručky sú z hľadiska obsahu a techniky spracovania zaradené onomastické príručky, resp. zdroje – Názvy obcí Slovenskej republiky (Vývin v rokoch 1773 – 1997) a Priezviská na Slovensku (podľa stavu v roku 1995), aktuálne sem pribudla ukážka Historického slovníka slovenčiny v podobe piateho zväzku (písmená R-rab – Š-švrkotať). Osobitné skupiny v položke Lingvistické zdroje tvoria elektronické verzie vybraných monografií a zborníkov z produkcie JÚĽŠ SAV alebo klasických autorov vrátane originálneho textu Štúrovej Nauky reči Slovenskej z r. 1846 a prvých Pravidiel slovenského pravopisu z r. 1931 a 1940, elektronické verzie štyroch zväzkov bibliografie Slovenských jazykovedcov (od r. 1925 do r. 2000) a posledných takmer pätnástich ročníkov troch časopisov vydávaných v JÚĽŠ SAV. Špecifickými položkami v lingvistických zdrojoch sú automatický prekladač do štúrovskej slovenčiny, ktorý vznikol ako ukážka možnosti automatického spracovania zmien v ortografii, a skúšobná verzia Slovenskej terminologickej databázy, ktorej cieľom je prispieť k zlepšovaniu odbornej komunikácie a zvyšovaniu terminologickej kultúry na Slovensku.

Okrem poslednej položky – Slovenskej terminologickej databázy, ktorá patrí medzi základné úlohy v rámci budovania Národného korpusu slovenského jazyka a elektronizácie jazykovedného výskumu na Slovensku, celý súbor lingvistických zdrojov sprístupnených na stránke JÚĽŠ SAV je akoby vedľajším produktom tohto projektu, ale veľmi dôležitým pre bežných používateľov slovenčiny, o čom svedčí aj priemerne 30-tisícová denná návštevnosť stránky JÚĽŠ SAV. On-line prístup k lingvistickým zdrojom veľmi oceňujú napr. lektori slovenského jazyka a iní záujemcovia o slovenčinu v zahraničí, ktorí inak nemajú veľa možností dostať sa k aktuálnym lexikografickým príručkám; www adresa JÚĽŠ SAV sa nachádza napr. aj na takej stránke, ako je http://how-to-learn-any-language.com – The website about teaching yourself languages.

Základnou úlohou skončeného i nadväzujúceho projektu, avšak nie s takým priamočiarym využitím, ako sme videli vyššie, je budovanie Slovenského národného korpusu v celej šírke štýlov, žánrov a vecných oblastí, ale aj regiónov, vydavateľstiev, generácií a pod., ktoré bolo v prvej fáze ohraničené na písané texty z obdobia rokov 1955 – 2006, v druhej fáze sa v spracúvaní písaných textov pokračuje tak, aby boli zachytené jednak tie najaktuálnejšie, jednak sa začína vytvárať samostatný celok textov spred roka 1955. Okrem písaných textov sa pripravuje aj korpus hovorenej slovenčiny, budujú sa paralelné korpusy a pod.

Slovenský národný korpus ako elektronický súbor jazykových dát s výkonnými nástrojmi na vyhľadávanie a triedenie skúmaných jazykových prostriedkov je od r. 2003 prístupný cez stránku JÚĽŠ SAV (v položke Lingvistické zdroje, ale aj v položke O ústave/Oddelenia), samostatne na adrese http://korpus.juls.savba.sk, neskôr bol odkaz naň zaradený aj na stránku Ministerstva školstva SR. Začiatkom r. 2007 bola sprístupnená najnovšia, šiesta verzia hlavného, základného korpusu prim-3.0, ktorá obsahuje 350 miliónov textových jednotiek (okrem slov je to aj interpunkcia a číslicové či iné neslovné zápisy ako emotikony a pod.), a spolu s ňou druhá verzia ručne morfologicky anotovaného korpusu v rozsahu takmer 512 tisíc textových jednotiek. Každý text v korpuse je podložený súhlasom autora alebo majiteľa autorských či distribučných práv na jeho spracovanie a zaradenie do celku korpusu podľa licenčnej zmluvy a má podrobnú bibliografickú a štýlovo-žánrovú anotáciu. Celý korpus je automaticky lematizovaný (každý slovný tvar má pri sebe informáciu o základnom tvare – leme) a automaticky morfologicky označkovaný po natrénovaní značkovacieho softvéru na ručne morfologicky anotovaných textoch. Vybrané texty sa ručne anotujú aj syntakticky. O postupoch pri získavaní textov, ako aj o princípoch ich spracovania od technického čistenia a konvertovania do jednotného formátu cez segmentáciu až po jednotlivé úrovne anotácie nájde záujemca podrobné informácie na stránke Slovenského národného korpusu.

Práca s korpusom je jednoduchá. Neregistrovaný používateľ má k dispozícii jednu, aktuálnu verziu veľkého korpusu prim a aktuálnu verziu ručne morfologicky anotovaného korpusu, s ktorými môže pracovať jednoduchým zadaním hľadaného slova alebo tvaru do vyhľadávacieho okienka priamo na stránke Slovenského národného korpusu. Voľne môže vyhľadávať aj v paralelných korpusoch a lingvistických zdrojoch. Registrovaní používatelia dostávajú na základe podpísania podmienok používania korpusu osobitný prístup s heslom na prácu s korpusom prostredníctvom korpusového manažéra Manatee s klientom Bonito a majú k dispozícii všetky verzie a podkorpusy (napr. samostatný podkorpus publicistických textov, samostatný podkorpus pôvodnej slovenskej umeleckej tvorby a pod.). V nich môžu nielen jednoducho vyhľadávať, ale uvedený počítačový nástroj im umožňuje vyhľadané kontexty (neraz desiatky tisícov dokladov) triediť, zisťovať ich rôzne štatistické hodnoty a distribúcie,

vytvárať kolokácie a pod. Vyselektovaný materiál si môžu uložiť do vlastného počítača a ďalej s ním pracovať. Všetci používatelia sú viazaní používať korpus výlučne na vedecko-výskumné a iné nekomerčné ciele a citovať korpus i jednotlivé zdroje v súlade s autorským zákonom.

Využívanie Slovenského národného korpusu (každoročne vyše 200 registrovaných používateľov, neregistrovaní sú zahrnutí v už spomínanej 30-tisícovej dennej návštevnosti) 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 v r. 2006 ukončený grant Morfosyntaktická analýza Slovenského národného korpusu (v spolupráci s FF PU Prešov) a zborník Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli (Prešov 2006), monografia M. Ivanovej Valencia statických slovies (Prešov 2006) či 2. zväzok Valenčného slovníka slovenských slovies (na korpusovom základe) autorky J. Nižníkovej (Prešov 2006), ako aj viacero jazykovedne zameraných diplomových prác. Najvýznamnejšou aplikačnou zložkou v tejto oblasti je lexikografické využitie: mnoho korpusov sa budovalo a buduje na podporu tvorby slovníkov a lexikografi sú v súčasnosti azda najčastejšími používateľmi korpusov – na báze materiálu Slovenského národného korpusu sa koncipuje nový 8-zväzkový Slovník súčasného slovenského jazyka a aktualizujú sa jednotlivé vydania doterajších lexikografických i pravopisných príručiek;

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 viacero diplomových a doktorandských prác na vysokých školách a výskumných pracoviskách technického zamerania, v záverečnom štádiu je vývoj vlastného morfologického analyzátora tvarov slovenského jazyka;

c) korpus môže byť a je dobrým zdrojom fráz a viet potrebných pri výučbe cudzieho, ale aj materinského jazyka, využívajú ho domáci učitelia i lektori slovenského jazyka v zahraničí na prípravu pravopisných, gramatických a štylistických cvičení; osobitne vítanou pomôckou v tejto oblasti sú paralelné korpusy, z ktorých sú v rámci Slovenského národného korpusu sprístupnené zatiaľ tri (Parallel Corpus of Computer Terms, Francúzsko-slovenský paralelný korpus, Rusko-slovenský paralelný korpus) a postupne sa plánujú ďalšie (najbližšie by to mal byť slovensko-český, slovensko-chorvátsky a slovensko-anglický paralelný korpus).

MÁRIA ŠIMKOVÁ

(Autorka je pracovníčkou Jazykového ústavu Ľudovíta Štúra SAV)