Na úžitok aj na parádu

■ Hovoríme s PhDr. Máriou Šimkovou, vedúcou oddelenia Slovenského národného korpusu JÚĽŠ SAV

Do povedomia verejnosti sa čoraz viac dostáva Slovenský národný korpus, fungujúci už päť rokov ako elektronická databáza slovenského jazyka, ktorá zahŕňa široké spektrum jazykových štýlov, žánrov i vecných oblastí a obsahuje prídavné jazykovedné informácie a výkonný vyhľadávací systém.

■ Pripomeňme si jeho podstatu, odvíjajúcu sa od pojmov korpus, korpusová lingvistika...

■ Keď sa pri prezentácii korpusu pýtame ľudí, čo si predstavujú pod týmto slovom, zvyčajne dostaneme odpoveď: piškótové cesto. A naozaj, ako v kuchyni, tak aj v lingvistike je korpus akýmsi základom, východiskovým lexikálnym materiálom, do ktorého sa ako plnka pridávajú lingvistické informácie (morfologické, syntaktické a pod.). Čerešničkou na torte je celá oblasť korpusovej lingvistiky, teda výskumov na veľkom množstve reálneho materiálu, a oblasť počítačového spracovania prirodzeného jazyka. Slovenský národný korpus je teda vedecko-výskumný projekt budovania elektronickej základnej slovnej zásoby, ktorý predstavuje špecifický súbor jazykových dát. Jeho základom sú texty zvyčajne rôznych štýlov, žánrov a vecných oblastí, ku ktorým sa pridávajú lingvistické informácie na úrovni slova, vety aj celého textu. Výkonné vyhľadávacie nástroje potom umožňujú vyhľadávanie a triedenie skúmaných jazykových prostriedkov a informácií. Na základe tohto autentického jazykového materiálu lingvisti opisujú významy a funkcie slov i ďalších jazykových javov. Najvýznamnejšou jazykovednou aplikačnou zložkou je lexikografické využitie: veľa korpusov sa budovalo a buduje na podporu tvorby slovníkov a lexikografi sú v súčasnosti azda najčastejšími používateľmi korpusov. Korpus však nenahrádza kodifikačné ani gramatické príručky, poskytuje „iba" materiál na ich prípravu.

Niektoré výsledky zo spracovania korpusov, ako sú zoznamy slov, spoločné výskyty slov, frekvencia slov atď., sa používajú aj v nelingvistických aplikáciách. Sem patria napríklad systémy na spracovanie textov (automatická kontrola pravopisu či gramatiky, strojový preklad textov) alebo systémy na rozpoznávanie reči. Korpus býva dobrým zdrojom príkladov potrebných pri výučbe slovenčiny ako cudzieho, ale aj materinského jazyka. Učebný počítačový program môže napríklad obsahovať klasický slovník spolu s menším korpusom, v ktorom sa dajú jednotlivé slová prezerať v kontexte, v akom sa reálne vyskytujú. Bežným používateľom jazyka môže korpus poslúžiť ako zdroj praktického poznania systému jazyka a overenia či doplnenia jednotlivých poznatkov.

■ Čomu všetkému sa venuje Slovenský národný korpus?

■ Slovenský národný korpus (SNK), tvorí iba 8 pracovníkov. To nie je ani jedna desatina v porovnaní s Českou republikou, kde sa počítačovou a korpusovou lingvistikou zaoberá približne sto ľudí na štyroch špecializovaných pracoviskách v Prahe i v Brne. V prvom rade je to budovanie korpusu písaných textov (v databáze SNK je aj Quark) a tvorba s tým súvisiacich počítačových nástrojov. Texty sa získavajú na báze licenčnej zmluvy s autormi alebo majiteľmi autorských či distribučných práv, v ktorej sa zaväzujeme využívať korpus výlučne na vedecko-výskumné a učebné ciele. A hoci sa korpus ako celok sprostredkúva používateľom cez internet, nemajú prístup k celým textom, ako je to v prípade elektronickej knižnice. Korpusový manažér im vždy poskytne iba určitý kontext (spravidla 100 slov), v ktorom sa nachádza hľadaný jazykový prostriedok. Takýchto kontextov môže byť niekoľko tisíc z rôznych diel. Každý text má presnú bibliografickú a štýlovo-žánrovú anotáciu, prostredníctvom ktorej sa použitý príklad dá citovať v súlade s autorským zákonom. Celý korpus je doplnený aj o základné lingvistické údaje – každému slovu je priradený základný tvar a informácia o morfologických kategóriách v danom kontexte. Používatelia vyhľadávajú v korpuse jazykové informácie pomocou korpusového manažéra Manatee a klienta Bonito z Fakulty informatiky Masarykovej univerzity v Brne. Môžu pracovať s veľkým korpusom v rozsahu okolo 350 miliónov slov, ktorý obsahuje všetky texty, alebo si môžu vybrať menší štýlovo vyvážený korpus či osobitné korpusy iba umeleckej, iba publicistickej alebo iba odbornej literatúry. K dispozícii je aj ručne morfologicky anotovaný korpus a paralelné korpusy, zatiaľ rusko-slovenský a francúzsko-slovenský, ale pripravujú sa už ďalšie: najbližšie chorvátsko-slovenský, česko-slovenský, nemecko-slovenský a anglicko-slovenský paralelný korpus. Tieto výstupy môžu poslúžiť najmä pri výučbe cudzieho jazyka, v zahraničí aj pri výučbe slovenčiny ako cudzieho jazyka, ale i v prekladateľskej praxi či opäť na porovnávacie výskumy.

Osobitnou, ale veľmi často navštevovanou položkou sú lingvistické zdroje a slovníky: tu sú používateľom bezplatne k dispozícii najnovšie kodifikačné príručky a rôzne publikácie z produkcie Jazykovedného ústavu Ľ. Štúra SAV alebo klasických autorov, napr. Štúrova *Nauka reči Slovenskej* v origináli.

■ V októbri sa v Bratislave zišli vedeckí a pedagogickí pracovníci z ôsmich krajín na medzinárodnej konferencii Slovko 2007, venovanej počítačovému spracovaniu prirodzeného jazyka, počítačovej lexikografii a terminológii. Konferenciu pripravilo vaše oddelenie a bolo to v poradí štvrté Slovko, čo už zakladá určitú tradíciu a zároveň svedčí o istom rešpekte zahraničných odborníkov k výsledkom hostiteľskej krajiny v spomenutej širšej oblasti.