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.

PhDr. Mária Šimková sa narodila v roku 1963 vo Vranove nad Topľou. Vyštudovala odbor slovenský jazyk a literatúra v kombinácii s históriou na prešovskej Filozofickej fakulte Univerzity P. J. Šafárika v Košiciach. Pred piatimi rokmi stála pri zrode oddelenia Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra Slovenskej akadémie vied. Ako vedúca tohto oddelenia bola zodpovednou riešiteľkou štátneho programu Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu (2003 – 2006). Venuje sa lexikologickej a gramatickej rovine jazyka, korpusovej lingvistike a celkovo využitiu korpusu. Pedagogicky pôsobila na viacerých vysokých školách.

Tieto konferencie sa konajú od roku 2001 pravidelne každý druhý rok v Bratislave. Prvé dva ročníky organizoval JÚĽŠ SAV spolu s Pedagogickou fakultou UK Bratislava, kde vtedy bolo partnerské pracovisko v podobe Laboratória počítačovej lingvistiky. Od r. 2005 organizuje Slovko oddelenie SNK JÚĽŠ SAV. Príspevky slovenských autorov na nedávnej štvrtej konferencii pokrývali základný aj aplikovaný výskum a poukazovali na prevratné zmeny, ktoré v ostatných rokoch zaznamenala naša počítačová a predovšetkým korpusová lingvistika.

Na konferencii odzneli najmä príspevky týkajúce sa slovanských jazykov. Súvisí to s tým, že aj slovenčina patrí medzi tieto jazyky a tak je úplne prirodzený častejší záujem a väčšia vôľa riešiť niektoré veci spoločne?

Počítačové spracovanie prirodzeného jazyka nie je úzko zamerané na slovanské jazyky, je to už niekoľko desaťročí celosvetovo aktuálna téma súvisiaca s informačnými a jazykovými technológiami. Praktické využitie vidíme všetci napríklad v tom, že máme pri písaní k dispozícii korektor, že môžeme na internete čítať informácie v slovenskom jazyku aj s diakritikou, čo ešte pred pár rokmi nebolo možné, že sa nám prostredníctvom internetu zobrazí to, čo hľadáme, teda že počítač akoby odpovedá na naše otázky, hoci niekedy sme z toľkých informácií aj unavení... Komunikácia s počítačom je dnes v mnohých oblastiach nevyhnutná, a tak je veľmi dobré, že môžeme aj v tejto špecifickej komunikácii používať slovenský jazyk.

Zameranie konferencie predovšetkým na slovanské jazyky súvisí nielen s tým, v akej oblasti sa nachádzame teritoriálne či jazykovo, ale aj s tým, v akej sme boli situácii, keď sa organizovala prvá či druhá konferencia s názvom Slovko: v roku 2001 ešte prakticky neexistovala elektronická databáza slovenských textov, hoci malý interný korpus slovenčiny sa v Jazykovednom ústave budoval a používal najmä na lexikografické účely. V roku 2002 sa podarilo získať osobitné financovanie korpusového projektu, začalo vznikať nové pracovisko Slovenského národného korpusu, takže pri prvých dvoch ročníkoch išlo najmä o získavanie skúseností z blízkych príbuzných (flektívnych) jazykov, ktorých počítačové spracovanie má svoje špecifiká a v niektorých okolitých krajinách bolo v tom čase na veľmi dobrej až špičkovej medzinárodnej úrovni (Česko, Poľsko, Slovinsko, Chorvátsko). Zatiaľ sa nám veľmi nedarí riešiť niektoré témy spoločne, pretože pri budovaní korpusu sa vždy objavia nejaké osobitosti konkrétneho jazyka alebo celkovo spoločnosti či národného spoločenstva, alebo sa neschvália konkrétne projekty (napr. WordNet, čo je elektronický sémantický slovník, do ktorého sme sa chceli zapojiť aj my so slovenským jazykom). Treba však zdôrazniť veľkú pomoc a podporu, ktorú sme vždy mali na pracoviskách v Českej republike (Praha, Brno), kde nám nezištne vychádzali v ústrety, poskytovali nielen dokumenty, literatúru, počítačové nástroje, ale sa najmä delili s vlastnými skúsenosťami a upozorňovali nás na rôzne prekážky, cez ktoré sa nám už potom ľahšie kráčalo.

Viete, že...

- **kolokačný slovník** obsahuje zvyčajne spájané dvojice alebo viac slov, samostatných lexikálnych jednotiek (napr. milé prekvapenie, veľký kus cesty).
- **retrográdny alebo obrátený slovník** má slová abecedne zoradené podľa poradia písmen od konca slova.
- **frekvenčný slovník** uvádza frekvenciu, t. j. početnosť výskytu slov, prípadne ich tvarov.

Skúste uviesť nejaký zaujímavý príklad, ktorý považujete za úspech vášho pracoviska v medzinárodnom meradle, teda taký, ktorý mohol osloviť aj zahraničných kolegov odborníkov.

V oblasti počítačového spracovania jazyka sa pozorne sleduje každý nový projekt, každý nový partner. Možno je to až neuveriteľné, ale nejde o likvidačnú konkurenciu – ide o spoluprácu, o spracovanie čo najväčšieho počtu jazykov, a každý jazyk si najlepšie spracujú domáci vedci. Kolegovia v Česku prejavujú o rozvoj Slovenského národného korpusu naozaj veľký záujem – čiže už naša existencia, existencia serióznej elektronickej databázy slovenského jazyka a systematická práca v oblasti korpusovej lingvistiky sa dostatočne oceňuje. A čo nám všetci úprimne závidia, to je sprístupňovanie slovníkov a ďalších lingvistických zdrojov širokej verejnosti na internete – to nemá ani jedna z okolitých krajín. Máme dobrú metodiku získavania textov s dôrazom na dodržiavanie autorských práv. Nezahanbíme sa ani v rovine morfologického značkovania textov. Aj paralelné korpusy poskytujú dostatok materiálu na výskum. Chystá sa Slovenský hovorený korpus, ktorým sa zapojíme do trendu budovania hovorených korpusov a výskumov v tejto oblasti.

V roku 2005 kolektív oddelenia Slovenského národného korpusu získal Cenu Slovenskej akadémie vied za budovanie infraštruktúry pre vedu, s čím súvisí aj budovanie terminologickej databázy. V akom štádiu je jej tvorba, čo všetko už zahŕňa a čo ešte bude obsahovať?

Slovenská terminologická databáza predstavuje jeden z našich najnovších príspevkov, hoci to nie je primárne korpusová záležitosť. No z odborných textov v korpuse sa postupne plánuje automatizovaný výber pojmov a súvislostí. Týmto projektom odpovedá Jazykovedný ústav Ľudovíta Štúra na požiadavku odborných kruhov a širokej verejnosti koordinovať vývoj jednotlivých terminológií a celkovú odbornú komunikáciu v slovenčine. Ide o voľne prístupnú databázu prostredníctvom internetu (***http://data.juls.savba.sk/std/***), ktorá bude centralizovať slovenské terminologické sústavy jednotlivých vedných disciplín a zároveň vytvárať priestor na spoluprácu a diskusiu s odborníkmi, čím prispeje k ustáleniu a zjednocovaniu sporných terminologických otázok. Skúšobná verzia databázy v súčasnosti obsahuje vyše 3 000 terminologických záznamov z 11 oblastí. Mala by sa primárne rozširovať o termíny z oblasti ekonómie a práva.

Chystajú sa ďalšie čerešne na korpusovú tortu v podobe kolokačného, retrográdneho a frekvenčného slovníka, hoci základné frekvenčné zoznamy sprístupňujeme na internete vždy s novou verziou korpusu. Tá najnovšia verzia príde symbolicky so začiatkom nového roka.

■ *Pripravila MIROSLAVA ČIERNA*
Snímka: KATARÍNA GAJDOŠOVÁ

http://korpus.juls.savba.sk/
http://slovnik.juls.savba.sk/
http://www.juls.savba.sk/