romboid5-11-text_romboid1-11-text.qxd 24.5.2011 15:26 Page 91

Z kuchyne Slovenského národného korpusu

Mária Šimková

Viacerí si ešte pamätáme nedávnu dobu *-izmov*, ktorú dnes, zdá sa, vystriedala doba *-ácií*: globalizácia, informatizácia, elektronizácia, digitalizácia... Všetka táto modernizácia neúprosne zasiahla aj slovenskú vedu, a to aj také (spoločenskovedné) odbory, kde sa donedávna vystačilo s hlavou, perom a kusom papiera.

Keď sa v roku 2002 na Slovensku zachytil vtedy už štyridsaťročný prúd korpusovej lingvistiky a v Jazykovednom ústave Ľ. Štúra SAV sa začal budovať Slovenský národný korpus, osemčlenný kolektív mladých ľudí, ktorí dovtedy pracovali v iných oblastiach lingvistiky, informatiky alebo aj jadrovej fyziky, začal objavovať skryté zákutia (nielen) nášho jazyka. Mohli sme sa pritom oprieť o kolegov z Českej republiky, ktorí sa už desaťročie venovali podobnej problematike a v tom čase dosahovali prvé výsledky porovnateľné so špičkovými korpusmi v Európe i vo svete. Na troch pracoviskách v Prahe a jednom v Brne (dohromady vyše sto zamestnancov [!]) sme stretávali nezištných radcov, ktorí nás nezabudli upozorniť i na vlastné omyly pri hľadaní najvhodnejších metód na počítačové spracovanie češtiny, aby uľahčili štart slovenského korpusu. Bol to nezabudnuteľný prejav kolegiality, s ktorou sme sa stretávali aj neskôr, keď sme dosiahli prvé cenné výsledky, a stretávame sa dodnes v spoločných projektoch, na konferenciách, v neformálnych kontaktoch.

Nebudeme tu však porovnávať naše jazyky, hoci aj na to by bolo dostatok zdrojov, ale sústredíme sa na slovenčinu a vrátime sa k začiatku a spôsobu budovania Slovenského národného korpusu. Predovšetkým: korpus ako elektronická databáza počítačovo a lingvisticky spracovaných textov, slúžiaca na vedecko-výskumné a učebné ciele, by nemohol vzniknúť a existovať bez textov. Ale odkiaľ a ako získať texty? Na základe zmluvy od ich tvorcov, spisovateľov, básnikov, autorov vedeckej spisby, prekladateľov, publicistov. Predpokladám, že viacerí čitatelia sa nachádzajú medzi týmito tvorcami a láskavými poskytovateľmi textov do korpusu, a hoci ich tu nemôžeme všetkých takmer šesťsto vymenovať, ďakujeme každému aspoň na našej internetovej stránke. Hneď v tejto základni korpusu – pri získavaní textov – máme viacero nástrah. Kým klasickí lingvisti mohli akékoľvek texty voľne excerpovať a analyzovať, v odbore korpusovej lingvistiky sa zhromažďuje do verejne prístupnej elektronickej databázy čo najväčšie množstvo textov v plnom rozsahu (celé texty kníh, novín, časopisov), aby mohli slúžiť jazykovedcom a iným bádateľom na analýzu akéhokoľvek jazykového prostriedku alebo javu, súboru jazykových prostriedkov, ich kombinácií, vzťahov, fungovania v rozličných typoch textov, v rozličných obdobiach, v písanej či hovorenej podobe jazyka. Na rozdiel od klasickej excerpcie sa zaradenie textu do korpusu môže uskutočniť až po podpísaní licenčnej zmluvy oprávňujúcej použiť texty na uvedené vedecko-výskumné ciele. Nuž, je to podmienka značne limitujúca, spôsobujúca, že mnohé diela, aj renomovaných autorov či vydavateľstiev, v Slovenskom národnom korpuse chýbajú a chýbajú z nich potom i autorské citáty v novom výkladovom Slovníku súčasného slovenského jazyka, ktorý čerpá dokladový materiál najmä z korpusu.

romboid5-11-text_romboid1-11-text.qxd 24.5.2011 15:26 Page 92

Kým neboli elektronické jazykové korpusy, lingvisti sa mohli pri svojich výskumoch oprieť zvyčajne len o malý a špecifický jazykový materiál, ktorý získali časovo náročnou vlastnoručnou excerpciou, a často sa museli spoliehať na vlastné jazykové vedomie, introspekciu a indukciu. V Jazykovednom ústave sa síce za desaťročia nahromadilo päť miliónov excerpčných lístočkov v tzv. všeobecnej súčasnej kartotéke, ale ich využitie bolo obmedzené na budovu ústavu a iba na heslové slová uvedené v záhlaví, podľa ktorých je celá „kamenná" kartotéka abecedne usporiadaná. Aby sa v kartotéke zachytilo čo najširšie spektrum jazykových jednotiek, nemohlo byť jedno slovo zaznamenané na príliš veľkom počte lístočkov. Z viac-menej jednotlivých dokladov potom nebolo možné zodpovedne preskúmať napr. spájateľnosť slova, jeho využitie v istých typoch textov či iné vlastnosti. Pri tvorbe korpusu (aj vďaka stále sa rozvíjajúcej počítačovej technike), naopak, platí zásada: čím viac textov, tým lepšie. Práve vtedy klesá entropia, ktorá sa pri malých rozsahoch nedá eliminovať ani najstarostlivejším výberom a štruktúrovaním textov, práve vtedy sa dá vyhodnotiť reálne fungovanie jazykových prostriedkov v ich jednotlivých významoch, tvaroch, spojeniach a pod.

Pravda, ak prostredníctvom Slovenského národného korpusu v súčasnosti ponúkame používateľom na jazykovú analýzu 700 miliónov jednotiek v základnom, klasickom korpuse (a ďalších vyše 950 miliónov v tzv. webovom korpuse), ľahko sa stane, že skúmané slovo či spojenie sa nájde v desiatkach, ba i v stovkách tisícov výskytov. A nie je v ľudských silách prejsť doklad po doklade a dôsledne analyzovať takéto množstvá kontextov ku každému sledovanému slovu. Na (pred)spracovanie sa preto využívajú počítačové nástroje, ktoré celý materiál usporiadajú podľa zadaných podmienok, vyselektujú „balast" nezodpovedajúci stanoveným kritériám, pripravia údaje o frekvencii, spájateľnosti, distribúcii v celom korpuse či v konkrétnych textoch. V Jazykovednom ústave sa na to využíva korpusový manažér Manatee s klientom Bonito z Fakulty informatiky v Brne a anglický softvér SketchEngine.

Pre počítačové nástroje je však zas potrebné lingvistické predspracovanie v podobe pridania jazykových informácií ku každému slovu (jeho základný tvar a morfologické kategórie), ku každej vete (aspoň jej začiatok a koniec), k celému textu (bibliografická a štýlovo-žánrová anotácia). Hoci texty v korpuse tvoria jednu veľkú, premiešanú masu, z ktorej sa používateľom „neservírujú" ucelené texty, ale len krátke kontexty hľadaných jazykových jednotiek postačujúce na jazykovú analýzu, predsa sa každé slovo musí dať morfosyntakticky identifikovať a zaradiť, kam autorsky i lingvisticky patrí. Jednotlivé úrovne spracovania pritom majú svoje nároky a pred človeka, ktorý potrebuje systém jazyka „vysvetliť" počítaču, kladú ďalšie a ďalšie nástrahy.

Čo je vlastne slovo? Lingvisticky ide o základnú pomenúvaciu jednotku bilaterálnej povahy, teda o jednotku s formou a obsahom. Počítač však obsahu nerozumie vôbec a formu rozoznáva ako reťazec znakov medzi dvoma medzerami. Takže napr. na oddelenie slova od interpunkcie, ktorá podľa pravopisných pravidiel nasleduje tesne za slovom bez medzery (okrem pomlčky), je v korpuse potrebné pridať pred interpunkčné znamienka umelé medzery. To hneď znamená, že pre počítač je „slovom" aj bodka, čiarka, výkričník a ďalšia interpunkcia, ale aj číslice, matematické a iné znaky a symboly. Na zabránenie terminologickej nejasnosti sa preto v korpusovej lingvistike považuje za základnú jednotku token, textová jednotka, zahŕňajúca slová, ale aj všetky znaky alebo súbory znakov oddelené medzerami.

Z morfologických kategórií si azda najskôr spomenieme na slovné druhy a na to, že slovenčina ich má desať. Ibaže počítač musí rozpoznať aj ostatné textové jednotky, a tak bolo potrebné vytvoriť pravidlá na rozlíšenie a označenie všetkých foriem v textoch. Okrem základných slovných druhov majú svoju triedu číslice, citátové slová, interpunkcia, skratky, slovnodruhovo neurčiteľné slová, neslovné znaky,

romboid5-11-text_romboid1-11-text.qxd 24.5.2011 15:26 Page 93

príčastia, polyfunkčné slová *sa, si* a slovo morfematickej povahy *by*. Spolu 19 tried. Niektoré sa dajú určiť ľahko, pri iných sa priradenie adekvátnej značky komplikuje homonymiou (*mať* je sloveso i podstatné meno, *heš* je citoslovce i priezvisko českého hokejistu, *dobre* je príslovka i tvar podstatného mena *dobro*), špecifickými vlastnosťami rôznych vlastných mien, novými a cudzími slovami s neštandardným zakončením atď. Na odstraňovanie nejednoznačnosti (dezambiguáciu) slúžia ďalšie výskumy, formalizácia ich výsledkov a implementácia týchto výsledkov do počítačových nástrojov.

Na základe lingvisticky predpripravených dát sa napríklad zisťuje frekvencia jazykových jednotiek. Na otázku, ktoré slovo v slovenčine je najfrekventovanejšie, s úsmevom odpovedáme: bodka a čiarka. Potom nasleduje pätica zložená z predložiek *v, na*, spojky *a*, pomocného slovesa *byť* a polyfunkčného *sa*. Prvým „ozajstným" slovom v našich korpusoch býva podstatné meno *rok*, ale to až okolo 30. miesta. V prvej stovke najfrekventovanejších jednotiek sú okrem ďalších predložiek a spojok časté ešte zámená a častice, pričom ide zväčša o jednopísmenové, jedno- alebo dvojslabičné slová. Táto prvá stovka predstavuje spolu vo všetkých výskytoch zhruba tretinu všetkých jednotiek v korpuse. Na otázku, koľko slov má slovenčina, pravdivo odpovedáme, že takýto špecifický výskum sme zatiaľ nerobili, ale z korpusových štatistík odhadujeme, že v najväčšej databáze je momentálne zachytených okolo 1,3 mil. jednotlivých (unikátnych výskytov) slov a asi tri milióny ich rôznych tvarov. A ktoré slovo je najmenej frekventované? To sa vypočítava ťažko, keďže s frekvenciou 1 a 2 sa v korpuse nachádza 1,7 mil. jednotiek typu znaky cudzích abecied, matematické znaky, nesprávne zapísané slová, kombinácie slov s číslicami alebo inými znakmi. Reálnych slov s výskytom 1 však môže byť aj 500 tisíc.

Väčšinu z týchto údajov možno nájsť na www stránke, kde sú, tak ako celý korpus a výsledky súvisiacich projektov (napr. Slovenská terminologická databáza), voľne prístupné na vedecko-výskumné a učebné využitie. Voľnosť však v tomto prípade neznamená, že sa z korpusu dá získať celý text nejakej knihy. Korpus nie je elektronická knižnica ani archív kníh, ale zdroj informácií o jazyku. Akýkoľvek používateľ, záujemca o slovenský jazyk môže hľadať konkrétne slovo či spojenie, aby si overil význam alebo pravopis – ale s vedomím, že korpus nie je kodifikačná príručka, keďže obsahuje texty v takej podobe, ako boli napísané a vydané, teda neraz i s chybami. Najčastejšími používateľmi sú samotní lingvisti (domáci aj zahraniční) a spomedzi nich najmä lexikografi, ktorí koncipujú slovníky. Významným výskumným zdrojom je korpus pre informatikov a matematikov (opäť aj zo zahraničia), s jeho využitím sa napísalo už viacero postupových prác. Špecifickými používateľmi sú psychológovia, neurológovia, logopédi, didaktici. V poslednom období „objavujú" možnosti využitia korpusu vo vyučovacom procese učitelia základných a stredných škôl. Ozaj, trúfli by ste si odhadnúť, či sa v textoch v korpuse vyskytuje častejšie *muž* alebo *žena*, *otec* alebo *matka*, *oko* alebo *zub*?

Mária Šimková pracuje v Jazykovednom ústave Ľudovíta Štúra SAV v Bratislave. Od roku 2002 vedie Slovenský národný korpus a je riešiteľkou viacerých projektov z oblasti korpusovej lingvistiky.