Možnosti a medze lingvistického výskumu v Slovenskom národnom korpuse[1]

Miloslava Sokolová – Mária Šimková – Martina Ivanová

1. Lingvistika zaznamenáva za posledných päťdesiat rokov nebývalú akceleráciu. Prejavuje sa to okrem iného v rozvoji nových, najmä tzv. hraničných disciplín (napr. psycholingvistika, sociolingvistika, etnolingvistika, filozofia jazyka) prevažne príbuzného – spoločenskovedného charakteru, ale pribudli aj odbory, ktoré zexaktňujú a zefektívňujú lingvistické výskumy – napr. matematická a počítačová lingvistika. V jej rámci sa už od 60. rokov minulého storočia rozvíja korpusová lingvistika napĺňajúc a súčasne sledujúc záujmy odborníkov z oblasti počítačového spracovania prirodzeného jazyka, spočiatku predovšetkým anglického. Pre jazykovedca-bádateľa to prinieslo nevyhnutnosť ďalšej (počítačovej) gramotnosti, postupné opúšťanie ručnej excerpcie, kartoték a takmer výlučnej opory v jazykovej introspekcii, ako aj množstvo otázok ohľadom nových metód výskumu a spracovania výsledkov z materiálu dovtedy nebývalého rozsahu. Od neraz odmietavo znejúcej otázky „načo korpus?“ a striktného odmietania dodatočného vnášania lingvistických informácií do „čistého“ súboru slov v reálnych kontextoch sa postupne prešlo k objavovaniu/otváraniu ďalších možností využitia korpusu na jazykový výskum a pristúpilo sa aj k lingvistickým (najmä morfologickým a syntaktickým) anotáciám korpusových textov. A to aj s vedomím existujúcich nástrah a rizík – anotácia, ani ručná, tobôž nie automatizovaná, nie je dokonalá a bezchybná, gramatické teórie sú rôzne a menia sa v čase atď. Do tohto procesu sa postupne zapojili aj jazyky flektívneho typu, ktorých formalizácia je síce náročnejšia, ale nie nemožná.

Slovensko zachytilo nástup budovania korpusov a rozvoja korpusovej lingvistiky v podstate až po r. 2000, keď už napríklad susední českí lingvisti mali k dispozícii 100-miliónový reprezentatívny korpus SYN2000. Najviac skúseností sa preto čerpalo aj vzhľadom na blízku príbuznosť jazykov práve z počítačovo- a korpusovolingvistických pracovísk v Českej republike.

[1] Názov sme aktualizovali podľa F. Štíchu, ktorý pomenoval svoju grantovú úlohu *Možnosti a meze gramatiky češtiny ve světle Českého národního korpusu*. Výskum v SNK prináša ďalšie obmedzenia (Sokolová, 2005).

2. Lingvistický výskum založený na korpusových zisteniach prináša bádateľovi nepopierateľne veľa výhod, jeho konečný efekt však závisí od kvality úrovne počítačového vybavenia pracovísk, ale aj od pripravenosti a schopnosti lingvistov/slovakistov tento potenciál využiť. Základnou výhodou je možnosť overenia a ilustrácie teoretických postulátov, ktorá sa už v súčasnom lingvistickom výskume považuje za conditio sine qua non, ako o tom hovorí F. Štícha: „... lingvistiku ‚deskriptivní´ či ‚popisnou´, jíž jde – zjednodušeně řečeno – o pravidla a/nebo pravidelnosti výskytu potenciálních entit jazykového systému v reálném diskursu (v textech, v promluvách, v ‚parole´), si nelze v budoucnu – dříve či později – představit jinak než jako lingvistiku ‚korpusovou´; minimálně v tom smyslu, že všechna její tvrzení o ‚existenci´, stylové hodnotě, frekvenci, textové distribuci či (míře) gramatičnosti uvažovaného výrazu budou vždy podložena korpusovými nálezy“ (Štícha, 2001, s. 161).

Korpusový materiál teda možno považovať za nevyhnutné empirické, deskriptívne východisko lingvistického výskumu, ktoré bude fungovať ako základ explanačného zhodnotenia. Takéto chápanie možno nájsť u F. Štíchu, ktorý navrhuje rozlišovať akceptabilitu/prijateľnosť jazykovej štruktúry, ktorá sa bude vzťahovať na reálne javy langovo-parolové, a gramatickosť, ktorá je vyhradená pre oblasť primárne langovú. Nulová potencialita jazykovej štruktúry (t. j. jej nulová realizácia v korpuse) znamená, že takúto štruktúru označíme ako negramatickú, pozitívna potencialita jazykovej štruktúry znamená jej doloženosť v korpuse, čo však nemusí značiť jej gramatickosť (porovnaj Štícha, 2001, s. 162). Napredujúce korpusovo-lingvistické výskumy (porov. napr. zborník *Korpus jako zdroj dat o češtině*, 2004) však ukazujú, že postulát o negramatickosti štruktúry vzhľadom na jej nulový výskyt v korpuse nemá reálnu platnosť. Na to by bol potrebný národný korpus jazyka, v ktorom by bol preukázateľne zachytený každý jazykový jav či prostriedok. Ak národné korpusy v súčasnosti disponujú len niekoľkými (dvoma, troma) stovkami miliónov textových jednotiek z (prevažne) písaných textov posledných desaťročí, nemôžeme to považovať za úplné zachytenie daného jazyka či reči. „... dostatečně velký korpus (sbírka dokladů parole) sice může být ‚objektivizací´ této parole, avšak z principiálních důvodů nemůže být ‚objektivizací´ langue... základním materiálem pro gramatický výzkum (tj. pro explikaci langue) je právě tato langue (jako implicitní znalost jazyka), korpus může být pro takovou práci pouze inspirací a korektivem“ (Oliva – Doležalová, 2004, s. 10). Korpusovú lingvistiku ako jazykovednú disciplínu, ktorá systematicky pracuje s korpusom a jeho nástrojmi, aby lepšie poznala funkcie a štruktúru jazyka, definoval aj F. Čermák (2000, s. 17).

3. Založením oddelenia Slovenského národného korpusu (ďalej SNK) v Jazykovednom ústave Ľ. Štúra SAV v roku 2002 sa otvorili nové možnosti aj v lingvistickom skúmaní súčasného slovenského jazyka. Hneď od roku 2003 sa

začala spolupráca oddelenia s FF PU v Prešove, na základe ktorej vznikol projekt *Morfosyntaktický výskum v rámci Slovenského národného korpusu* (MŠ SR VEGA 1/3149/04).

Kolektív SNK bol v r. 2002 inštitucionalizovaný, aby realizoval *projekt vybudovania Národného korpusu slovenského jazyka a projekt elektronizácie jazykovedného výskumu v rokoch 2002 – 2006*. Od roku 2003 do roku 2006 kolektív súčasne riešil úlohu výskumu a vývoja *Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu* tematického štátneho programu výskumu a vývoja *Aktuálne otázky rozvoja spoločnosti*. Riešitelia zvolili nie celkom zvyčajný, pre používateľov aj trochu náročnejší postup sprístupňovania korpusu verejnosti (http://korpus.juls.savba.sk). Nečakalo sa totiž na zhromaždenie dostatočného východiskového materiálu, z ktorého by sa po čase dal vytvoriť vyvážený či reprezentatívny korpus relevantnej veľkosti (odhaduje sa, že treba mať k dispozícii aspoň päťnásobné množstvo textov, aby sa dal „namixovať" korpus vyvážený, reprezentatívny nielen z hľadiska štýlov, ale aj žánrov, regiónov, vedných odborov, umeleckých škôl, období a pod.), no priebežne sa sprístupňovali všetky texty. S každým významnejším prírastkom a zlepšením anotácie sa zverejňovala nová verzia, niekedy to boli aj dve verzie v priebehu roka. Od 30-miliónového korpusu zloženého takmer úplne z publicistických textov vo verzii prim-0.1 v r. 2003 sa SNK do konca r. 2006 rozrástol na približne 350 miliónov textových jednotiek s približne 20-percentným podielom umeleckých a 20-percentným podielom odborných textov vo verzii prim-3.0. Bol to pragmatický postup vzhľadom na dlhodobú absenciu korpusu slovenského jazyka a veľké zaostávanie za okolitými krajinami, ako aj vzhľadom na potreby koncipovania nového výkladového *Slovníka súčasného slovenského jazyka*. Na lingvistický výskum v rámci grantu to však kládlo zvýšené nároky pri udržiavaní konzistencie najmä pri zisťovaní a hodnotení frekvencie skúmaných javov.

Rovnako náročné, spočiatku až problematické bolo využívanie morfologickej anotácie. V r. 2004 a 2005 bola k dispozícii len automatizovaná anotácia pomocou českého softvéru a českých značiek, ktorá mohla byť nápomocnou a orientačnou iba pri sledovaní niektorých javov, napr. základných kategórií pri substantívach a verbách s výnimkou vidu, ale na výskum napr. neflektívnych slovných druhov nebola takmer vôbec vhodná vzhľadom na rozdielnosť gramatických teórií a rozdielnych slovnodruhových charakteristík v slovenskej a českej lingvistike. Ručná morfologická anotácia pomocou slovenských pravidiel a značiek (Garabík – Gianitsová – Horák – Šimková, 2004) napredovala pomaly, prvý ručne anotovaný korpus bol sprístupnený až začiatkom r. 2006, aj to iba v rozsahu 322 600 textových jednotiek. Automatizovaná anotácia slovenskými značkami na jeho základe bola značne nespoľahlivá. Zdokonalenie morfologickej anotácie na konci r. 2006 (ručne anotovaných 511 534 textových

jednotiek, automatizovaná anotácia vylepšená aj zapojením morfologického analyzátora a generátora tvarov slovenského jazyka vyvinutého v oddelení SNK) už bolo mimo možnosti zahrnutia do výskumu v tomto grante.

4. Keďže riešitelia grantu VEGA nemali do decembra 2005 k dispozícii verziu štýlovo vyváženého, kvalitne anotovaného korpusu, pôvodné ciele výskumu sa museli modifikovať, porov. aj názov tohto zborníka *Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli*. Zameranie na výskum vybraných verbálnych a substantívnych tvarov a lexém zväčša korešpondovalo s dlhodobou vedecko-výskumnou orientáciou členov riešiteľského kolektívu a dalo sa realizovať aj v podmienkach budovania a premenlivosti materiálovej bázy. Spracovanie vybraných jazykových javov je v štúdiách zborníka prezentované zámerne v podobnej štruktúre ako v *Longman Grammar of Spoken and Written English* (1999) – v prehľadných tabuľkách a grafoch podľa jednotnej úpravy. Na rozdiel od uvedenej gramatiky v kapitolách zo slovenskej korpusovej morfosyntaxe sa vzhľadom na chýbajúci korpus hovorených prejavov muselo vychádzať len z písaných textov z troch štýlových oblastí bez konverzačných (hovorených) textov. Korpus ako východisko pre morfosyntaktický výskum bol zúžený na publicistické (centrálne a regionálne) texty, odborné, populárno-vedecké a umelecké texty v adekvátnom pomere, čo znamená na rozdiel od existujúcich gramatík výrazný posun k publicistickým a vedeckým textom (vo vyváženom korpuse prim-2.1-vyv bol pomer 60 % publicistická literatúra, 20 % umelecká literatúra, 20 % odborná literatúra ako odraz zastúpenia štýlov v súčasnej produkcii aj percepcii textov na Slovensku). V tomto zmysle vyvážený elektronický korpus (externe anotovaný – štýl, žáner) je výberom zo všetkých textov v *Slovenskom národnom korpuse* (Šimková, 2004). Zatiaľ sa nám v lingvistickom výskume potvrdzuje, že čím väčšia starostlivosť sa venuje selekcii východiskových textov, tým objektívnejšie výsledky možno získať aj z početne menšieho korpusu. Neplatí to však pre všetky jazykové javy rovnako.

Autori štúdií pracovali podľa jednotnej metodiky pomocou korpusového manažéra Manatee s klientom Bonito. Spôsob, ako z korpusu získať požadované lingvistické informácie, je vyhľadávanie slova alebo tvaru v kontextových použitiach pomocou špeciálneho konkordančného (vyhľadávacieho) programu. Konkordančný program jednotlivé javy nielen zobrazí tak, ako sa v konkrétnych kontextoch nachádzajú, ale poskytuje o nich aj frekvenčné, štatistické a bibliografické údaje, napr. absolútny počet výskytov v danom korpuse, výskyt v jednotlivých štýloch, výber autora, titulu a štýlu textu. Program umožňuje skúmať aj širší kontext dokladov, ako je prednastavený počet slov pred alebo za hľadaným výrazom. Z ďalších nástrojov je nevyhnutné využívať filtre: pozitívny filter pri veľkom množstve javov rozličnej hierarchie, negatívny filter pri redundantných informáciách. Keďže vo väčšine prípadov sme skúmali tvary,

okrem hľadania pomocou lemy sme vyhľadávali aj konkrétne tvary pomocou funkcie word. Vyhľadávanie pomocou tagov bolo dosť nespoľahlivé, preto veľa času zabrala individuálna ručná kontrola materiálu. Predovšetkým pri výskume valencie je dôležitá štatistika kolokácií.

Doterajší výskum v troch verziách *Slovenského národného korpusu* potvrdzuje, že kým na výskum frekventovaných jazykových prostriedkov lepšie vyhovuje frekvenčná distribúcia vo vyváženom korpuse (prim-2.0-vyv, prim-2.1-vyv), okrajové jazykové prostriedky treba skúmať v celých korpusových verziách (prim-2.0-all, prim-2.1-all). Keďže, ako sme už uviedli, na začiatku výskumu nebola k dispozícii verzia štýlovo vyváženého, spoľahlivo anotovaného korpusu, začali sme skúmať okrajové javy slovenského gramatického systému, ktorých nižšia frekvencia dovoľovala aj manuálnu kontrolu získaného materiálu (Stašková – Sokolová – Kášová, 2005).

Silnou stránkou korpusových výskumov sú údaje o frekvencii dokladov, preto sa v každej štúdii uvádza:

- frekvencia morfosyntaktických prostriedkov v korpuse (celkovo podľa štýlov – štýlová disperzia);
- pomerná frekvencia (v pomere k podielu daných textov), ktorá má vyššiu výpovednú hodnotu než celková frekvencia, porov. Stašková – Sokolová – Kášová, 2005.

Pri interpretácii morfosyntaktických javov sa využíva možnosť, ktorú korpusový výskum ponúka – overenie komunikačných funkcií v prirodzených komunikačných modeloch (porov. štúdie o kondicionáli). Pri každom skúmanom jave sme interpretovali relevantné parametre jeho statusu vyplývajúce z jeho východiskovej definície.

5. Osobitne cenným prínosom práce na grante *Morfosyntaktická analýza Slovenského národného korpusu* je spolupráca študentov a doktorandov. V oddelení SNK JÚĽŠ SAV sa na lingvistických anotáciách podieľali študenti a absolventi Filozofickej fakulty Prešovskej univerzity v Prešove, Filozofickej fakulty Katolíckej univerzity v Ružomberku, Univerzity Cyrila a Metoda v Trnave, Filozofickej fakulty Univerzity Komenského v Bratislave. Viacerí z nich sa zapojili aj do prípravy materiálu a spracovania výsledkov výskumu v tomto zborníku. Výsledkom doterajšieho zapojenia študentov FF PU do výskumu sú aj diplomové práce pod vedením M. Sokolovej: výskum transgresívu a činného prítomného particípia (Zdena Bugajová), výskum verbálneho substantíva (Erika Gregusová), výskum imperatívu (Anna Krellová), výskum pronomín (Júlia Timčová), výskum numerálií (Daniela Kalaninová), výskum kondicionálu (Vladimír Dziak a Slavomíra Rabatinová), a pod vedením J. Nižníkovej: výskum valencie slovies (Zuzana Mattová, Daniela Vidová). Viacerí študenti a doktorandi vystúpili s čiastkovými výsledkami na vedeckých konferenciách.

Pri získavaní inšpirácií, zadávaní a vedení diplomových a doktorandských prác bola dôležitá tímová spolupráca sústredená do jedného týždňa v rámci viacerých špeciálnych workshopov zvyčajne pred začiatkom semestra. Na workshopoch sa zúčastňovali aj predstavitelia z českých pracovísk, napr. František Štícha (Ústav pro jazyk český AV ČR Praha), Jan Hajič, Zdeňka Urešová a Veronika Kolářová (Ústav formální a aplikované lingvistiky MFF UK Praha), od ktorých získali riešitelia projektu významnú metodickú pomoc.

Práca na grante *Morfosyntaktická analýza Slovenského národného korpusu* sa napokon stala prvou koncepčnou a systematickou sondou do možností a obmedzení korpusového výskumu na Slovensku. Pri uzatváraní a rekapitulácii tejto prvej fázy môžeme konštatovať, že bola vytvorená základná metodika lingvistickej práce s rozsiahlym korpusom textov, získali sa štatisticky presné informácie o distribúcii skúmaných javov, čo môže prispieť k ich presnejšiemu opisu v budúcich pravopisných či gramatických príručkách. Keďže sa medzitým zväčšila a skvalitnila materiálová báza *Slovenského národného korpusu*, veríme, že podobné výskumy budú pokračovať a prinášať ďalšie relevantné výsledky, z ktorých bude možné vytvoriť syntetický morfosyntaktický opis slovenského jazyka.

Literatúra

BIBER, D. – JOHANSSON, S. – LEECH, G. – CONRAD, S. – FINEGAN, E.: Longman Grammar of Spoken and Written English. Harlow: Longman 1999.

BUGAJOVÁ, Z.: Deklinácia cudzích proprií. In: Študentská vedecká konferencia. Zborník abstraktov. Zost. D. Slančová – I. Žarnovská. Prešov: Filozofická fakulta Prešovskej univerzity 2005, s. 11 – 12.

CZABALOVÁ, L.: Analýza aj, ani, i – využitie a fungovanie v Slovenskom národnom korpuse. In: Študentská vedecká konferencia. Zborník abstraktov. Zost. D. Slančová – I. Žarnovská. Prešov: Filozofická fakulta Prešovskej univerzity 2006, s. 11 – 12.

ČERMÁK, F.: Jazykový korpus: Prostředek a zdroj poznání. In: Studie z korpusové lingvistiky. Acta Universitatis Carolinae. Philologica 3 – 4. Praha: Univerzita Karlova – Nakladatelství Karolinum 2000, s. 15 – 37.

DZIAK, V. – RABATINOVÁ, S.: Korpus verzus tradícia. In: Študentská vedecká konferencia. Zborník abstraktov. Zost. D. Slančová – I. Žarnovská. Prešov: Filozofická fakulta Prešovskej univerzity 2005, s. 17 – 19.

DZIAK, V. – RABATINOVÁ, S.: Temporálna súslednosť na základe korpusových zistení. In: Študentská vedecká konferencia. Zborník abstraktov. Zost. D. Slančová – I. Žarnovská. Prešov: Filozofická fakulta Prešovskej univerzity 2006, s. 17 – 19.

GARABÍK, R. – GIANITSOVÁ, L. – HORÁK, A. – ŠIMKOVÁ, M.: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. In: http://korpus.juls.savba.sk, 2004.

GIANITSOVÁ, L.: Zamyslenie nad výučbou zámen a čísloviek pri príprave morfologickej anotácie SNK. In: Tradiční a netradiční metody a formy práce ve výuce českého jazyka na základní škole. Sborník prací z mezinárodní konference konané 19. 3. 2004 na Pedagogické fakultě UP v Olomouci. Ed. M. Polák, K. Vodrážková. Olomouc: Univerzita Palackého, 2005, s. 53 – 65.

GIANITSOVÁ, L.: Jazykový korpus a nové dimenzie výskumu a výučby jazyka. In: Slovo o slove. Zborník Katedry slovenského jazyka a literatúry PdF Prešovskej univerzity. Roč. 11. Prešov: KSJL PdF PU, 2005, s. 27 – 36.

GIANITSOVÁ, L.: K morfologickej variantnosti v deklinácii feminín a neutier (niektoré frekvenčné ukazovatele). In: Študentská vedecká konferencia. Zborník abstraktov. Zost. D. Slančová – I. Žarnovská. Prešov: FF PU 2005, s. 20 – 22.

HORÁK, A. – GIANITSOVÁ, L. – ŠIMKOVÁ, M. – ŠMOTLÁK, M. – GARABÍK, R.: Slovak National Corpus. In: Text, Speech and Dialogue. 7th International Conference TSD 2004 Proceedings. Ed. P. Sojka, I. Kopeček, K. Pala. Berlin – Heidelberg: Springer – Verlag 2004, s. 89 – 94.

Korpus jako zdroj dat o češtině. Ed. P. Karlík. Brno: Masarykova univerzita v Brně 2004.

LOPATKOVÁ, M. – ŽABOKRTSKÝ, Z. a kol.: Tektogramaticky anotovaný valenční slovník českých sloves. Praha: Universitas Carolina Pragensis 2002.

OLIVA, K. – DOLEŽALOVÁ, D.: O korpusu jako o zdroji jazykových dat. In: Korpus jako zdroj dat o češtině. Ed. P. Karlík. Brno: Masarykova univerzita v Brně 2004, s. 7 – 10.

Slovenčina a čeština v počítačovom spracovaní. Zborník referátov zo seminára (Bratislava 26. – 27. októbra 2001). Red. A. Jarošová. Bratislava: Veda 2001.

Slovenský národný korpus. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2005. Dostupný z WWW: http://korpus.juls.savba.sk.

Slovenský národný korpus. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2006. Dostupný z WWW: http://korpus.juls.savba.sk.

SOKOLOVÁ, M.: Koncepcia slovenskej korpusovej morfosyntaxe. In: Jazyk a komunikácia v súvislostiach. Bratislava: FF UK, 2005 s. 284 – 297.

STAŠKOVÁ, J. – SOKOLOVÁ, M. – KÁŠOVÁ, M.: Verifikácia tvrdení v tradičných gramatikách korpusom (SNK) – okrajové jazykové prostriedky. In: Gramatika & Korpus. Praha: Ústav pro jazyk český Akademie věd ČR 2005, s. 60 – 61.

ŠIMKOVÁ, M.: Možnosti využitia SNK na štúdium slovenského jazyka. In: Studia Academica Slovaca. 33. Red. J. Mlacek, Bratislava: Stimul – Centrum informatiky a vzdelávania FF UK 2004, s. 204 – 217.

ŠTÍCHA, F.: Gramatický výskum dříve a dnes: korpus jako výzva. In: Tradícia a perspektívy gramatického výskumu na Slovensku. Zost. M. Šimková. Bratislava: Veda 2003a. s. 24 – 31.

ŠTÍCHA, F.: Česko-německá srovnávací gramatika. Praha: Argo 2003b.

ŠTÍCHA, F.: Kritéria gramatičnosti (Korpus jako argument a inspirace). In: Slovo a slovesnost, 62, 2001, s. 161 – 175.

ŠTÍCHA, F. – HOLUBOVÁ, V.: Okazionální slovesné vazby v českém národním korpusu. In: Jazyky a jazykověda. Praha: Filozofická fakulta Univerzity Karlovy – Ústav českého národního korpusu 2004, s. 273 – 284.

Teoretické základy synchronní mluvnice spisovné češtiny. In: Slovo a slovesnost, 1975, s. 18 – 46.
Tradícia a perspektívy gramatického výskumu na Slovensku. Zost. M. Šimková. Bratislava, Veda 2003.