EDÍCIA ŠTÚDIE Z KORPUSOVEJ LINGVISTIKY A JEJ PRVÝ ZVÄZOK

(Korpusová lingvistika: Stav a modelové přístupy. Studie z korpusové lingvistiky. Sv. 1. Ed.: F. Čermák – R. Blatná. Praha: Nakladatelství Lidové noviny 2006. 358 s. ISBN 80-7106-865-9)

Mária Šimková – Katarína Gajdošová
Slovenský národný korpus Jazykovedného ústavu Ľudovíta Štúra SAV
Panská 26, 813 64 Bratislava
e-maily: marias@korpus.sk, katarinag@korpus.sk

Na úvod krátka história

Ústav Českého národného korpusu vznikol pod vedením Františka Čermáka v r. 1994 na Filozofickej fakulte Univerzity Karlovej v Prahe s cieľom vybudovať reprezentatívnu elektronickú databázu českého jazyka v jeho rôznych podobách. Keď sa z tohto pracoviska v r. 2000 sprístupnil na internete prvý 100-miliónový korpus češtiny SYN2000, záujemcovia o prácu s ním dostali k dispozícii aj praktickú príručku (Český národní korpus – úvod a příručka uživatele, 2000), no najmä prehľad o hlavných odvetviach a smeroch korpusovej lingvistiky v podobe výberového súboru 22 štúdií od popredných svetových odborníkov v tejto oblasti. Išlo o obsiahlu knihu prekladov predovšetkým metodologicky a kriteriálne orientovaných príspevkov, ktorú jej editori F. Čermák, J. Klímová a V. Petkevič nazvali ***Studie z korpusové lingvistiky*** (2000).

Snahou vydavateľov a prekladateľov bolo predstaviť korpusovú lingvistiku nielen ako zdroj a sprostredkovateľa pre iné odbory, ale aj ako samostatný odbor s vlastnou rozvíjajúcou sa terminológiou (súčasťou publikácie je i anglicko-český slovník odborných termínov), s vlastným predmetom a cieľmi. Kniha bola zároveň zamýšľaná ako zdroj inšpirácie pre podobnú prácu s Českým národným korpusom (ďalej ČNK). Tento zámer sa postupne začal napĺňať jednak vďaka tvorbe a sprístupňovaniu ďalších korpusov (nielen v rámci ČNK, ale aj v Ústave formálnej a aplikovanej lingvistiky Matematicko-fyzikálnej fakulty UK v Prahe a na Katedre informačných technológií Fakulty informatiky Masarykovej univerzity v Brne, ktoré ako pracoviská technického charakteru vyvíjajú aj príslušné softvérové nástroje), jednak vďaka záujmu lingvistov, ktorí pochopili význam korpusovej metódy a využitia materiálu nebývalého rozsahu na prehĺbenejšie poznanie jazyka. Na nie jednoduché vyrovnávanie sa s týmito novými možnosťami upozornil napr. projekt F. Štíchu s príznačným názvom Možnosti a meze gramatiky češtiny ve světle Českého národního korpusu a viacero jeho čiastkových výstupov: Korpus jako zdroj dat o češtině (2004), Gramatika a korpus / Grammar & Corpora 2005 (2007), Možnosti a meze české gramatiky (2006).

Edícia Studie z korpusové lingvistiky (2006 – ...)

V predhovore k prvému zväzku hlavný editor F. Čermák konštatuje, že sa ním začína séria, ktorej cieľom je zverejňovanie najdôležitejších výsledkov práce lingvistov s korpusom. „Tato řada navazuje na stejnojmennou knihu ..., která v podobě základní oborové čítanky představila v překladu některé z nejvýznamnějších prvních příspěvků v oboru korpusová lingvistika. Nicméně práce s Českým národním korpusem se překvapivě rozmohla natolik, že už je tu nová řada i domácích autorů, jejichž výsledky a analýzy stojí za to zveřejnit a naznačit tak pokračování po předchozím „nultém" svazku" (s. 7). Jednotlivé zväzky majú mať podľa vyjadrenia hlavného editora pracovnú povahu a majú predstavovať reálne a aktuálne výsledky práce s korpusom.

Od r. 2006 obsahuje edícia Štúdie z korpusovej lingvistiky už 13 zväzkov v podobe zborníkov i monografií:

2. Kolokace. Ed.: F. Čermák – M. Šulc. Praha: Nakladatelství Lidové noviny 2006. 452 s.
3. Blatná, Renata: Víceslovné předložky v současné češtině. Praha: Nakladatelství Lidové noviny 2006. 351 s.
4. Kopřivová, Marie: Valence českých adjektiv. Praha: Nakladatelství Lidové noviny 2006. 125 s.
5. Esvan, François: Vidová morfologie českého slovesa. Praha: Nakladatelství Lidové noviny 2007. 342 s.
6. Šonková, Jitka: Morfologie mluvené češtiny: Frekvenční analýza. Praha: Nakladatelství Lidové noviny 2008. 356 s.
7. Čeština v mluveném korpusu. Ed.: M. Kopřivová – M. Waclawičová. Praha: Nakladatelství Lidové noviny 2008. 273 s.
8. Cvrček, Václav: Regulace jazyka a Koncept minimální intervence. Praha: Nakladatelství Lidové noviny 2008. 230 s.
9. Čermáková, Anna: Valence českých substantiv. Praha: Nakladatelství Lidové noviny 2009. 124 s.
10. Šimandl, Josef: Dnešní skloňování substantiv typů *kámen, břímě*. Praha: Nakladatelství Lidové noviny 2010. 374 s.
11. Načeva-Marvanová Mira: Perfektum v současné češtině. Praha: Nakladatelství Lidové noviny 2010. 224 s.
12. Mnohojazyčný korpus InterCorp: Možnosti studia. Ed.: F. Čermák – J. Kocek. Praha: Nakladatelství Lidové noviny 2010. 292 s.
13. InterCorp: Exploring a Multilingual Corpus. Ed.: F. Čermák – A. Klégr – P. Corness. Praha: Nakladatelství Lidové noviny 2010. 253 s.

Mnohé z predstavených výskumov sa realizujú aj na Slovensku na materiáli Slovenského národného korpusu a špecializovaných podkorpusov, takže dosiahnuté výsledky možno priamo porovnávať a konfrontovať slovenčinu a češtinu ako dva

blízkopríbuzné jazyky. Viaceré témy však u nás čakajú na podobné monografické spracovanie.

Čitateľov Slovenskej reči budeme o jednotlivých zväzkoch Štúdií z korpusovej lingvistiky postupne informovať v sérii recenzií a rozhľadových článkov.

Korpusová lingvistika: Stav a modelové přístupy. Studie z korpusové lingvistiky. Sv. 1. Ed.: F. Čermák – R. Blatná. Praha: Nakladatelství Lidové noviny 2006. 358 s.

18 príspevkov je síce v zborníku zoradených abecedne podľa priezvísk autorov, ktorých je 17, ale v našom prehľade sa ich pokúsime usporiadať tematicky. Viaceré štúdie sú metodologického a kriteriálneho charakteru, čím tento prvý zväzok edície priamo nadväzuje na spomínaný východiskový súbor prekladov.

Úvod zborníka predstavuje zhrnujúca štúdia F. Čermáka *Korpusová lingvistika dnešní doby* (s. 9 – 18). Na jednej strane autor poskytuje ešte východiskový opis odboru vrátane základných termínov (*korpus, korpusová lingvistika, anotácia, konkordancia, korpusový manažér, kolokácia, lema*)[1], na druhej strane zužitkúva niekoľkoročné skúsenosti z práce s korpusmi a vystihuje ich prínosy a odlišnosti od iných zdrojov (jazykových) informácií (napr. internet), ich nedostatky, ako aj nové tendencie v ich využívaní. Z prínosov je to okrem rozsahu a pestrosti dát (korpusy písaných textov s miliardami jednotiek, historické, hovorené, paralelné a i. korpusy) najmä možnosť využitia korpusových textov na skúmanie súvislostí, spoluvýskytov a textových súvzťažností, teda syntagmatiky jazyka, na rozdiel od predchádzajúceho takmer výlučne paradigmatického prístupu a sledovania viac-menej izolovaných javov. Nedostatky korpusov a korpusovej metodológie, ako na nich upozorňujú aj autori ďalších príspevkov, spočívajú paradoxne opäť v rozsahu a pestrosti jazykových dát, ktoré prinášajú nemalé formálne, technické aj lingvistické problémy. Ide napr. o spoľahlivosť anotácie, potrebu nástrojov na triedenie a selekciu ľudsky často nezvládnuteľného rozsahu materiálu, variabilnosť jazykových prostriedkov, homonymiu, rozpoznanie viacslovných jednotiek. A ako ukazuje Jan Králík vo svojom *Zamyšlení nad velkými výběry* (s. 205 – 209) z pohľadu štatistiky, veľké textové korpusy nespôsobili isté problémy iba v lingvistike, no zasiahli aj oblasť matematickej teórie, jej aplikačných postupov, ale aj základných stĺpov teórie pravdepodobnosti, predovšetkým axióm a zákona veľkých čísel. K výsledkom získaným z korpusov treba vo všeobecnosti pristupovať vo svetle obmedzenosti vs. neobmedzenosti dát (ak niečo nie je v korpuse, to ešte neznamená, že to nie je v jazyku) a ich matematicko-štatistického videnia (porovnávanie porovnateľných dát a hodnôt).

[1] Samostatný širší súpis pojmov z korpusovej lingvistiky v češtine bol uverejnený napr. v publikácii M. Šulca (1999); podobný výberový terminologický slovníček sme vypracovali aj pre slovenčinu (Šimková, 2006) a zaradili ho aj do Slovenskej terminologickej databázy (https://data.juls.savba.sk/std/).

Ako v matematike, tak aj v lingvistike sa dá na korpusy pozerať ako na narušitele osvedčených postupov, spochybňovatele dosiahnutých výsledkov a dlhé roky vypracúvaných teórií alebo – pokúsiť sa korpusy využiť na overenie či precizovanie doterajších poznatkov. Viacerí autori prezentujú pozitíva (i obmedzenia) poslednej možnosti dokonca pri skúmaní syntaktických javov.

Eva Hajičová zdôraznila, že lingvisticky (t. j. aj syntakticky) anotované dáta sú „nedocenitelným testem pro lingvistickou teorii stejně jako dosud nevídaným zdrojem informací o daném jazyce, které je možné s výhodou využít pro budování nových gramatik“ (s. 118). V príspevku *Využití korpusu pro ověřovaní lingvistických hypotéz* (s. 118 – 130) rozvinula niektoré svoje predchádzajúce hypotézy a pozorovania a zhrnula aj čiastkové výsledky ďalších analýz rozpracovaných na túto tému. Overované hypotézy sa týkajú aktuálneho členenia vety[2], vlastností syntaktickej štruktúry a analýzy diskurzu. Napr. v hypotéze B autorka formulovala predpoklad, že závislostné stromy na tektogramatickej (hĺbkovej) rovine sú projektívne, t. j. nedochádza k pretínaniu vetiev smerujúcich od riadiacich k závislým členom. Analýzou zistené odchýlky od tejto podmienky predstavujú syntaktické štruktúry s predsunutými zámenami, časticami, pomocnými slovesami a pod., ale aj rozdelené syntagmy v prípadoch uplatnenia základného slovosledného princípu českej vety – umiestnenia kontrastu v počiatočnej pozícii (***Společnou*** *máme především tuto* ***odpovědnost***.). Uvedený typ predstavuje v súbore viet Pražského závislostného korpusu 6 % zo zistených odchýlok.

V ďalších troch príspevkoch sa autori venovali nielen preskúmavaniu možností využitia korpusu na overenie hypotéz lingvistickej teórie, ale aj formulovaniu pravidiel na zlepšenie anotácie korpusu a získaniu podkladov na budovanie formálnej gramatiky. Vladimír Petkevič sa v prípadovej štúdii ako príspevku k automatickej dezambiguácii českých textov zameral na *Automatické rozpoznání infinitivu* (s. 226 – 253) a Alexander Rosen doplnil, *O čem vypovídá pád doplňku infinitivu* (s. 254 – 284). Petkevičov súpis homonymných infinitívnych a iných slovných tvarov tvorí 6 skupín, z ktorých dezambiguačne najnáročnejšou a zároveň najfrekventovanejšou je skupina obsahujúca tvary *stát, moci, pomoci, růst, nemoci, obrat, drát, vzrůst, srůst*.[3] Autor analyzuje všetky syntaktické vlastnosti slovesných významov uvedených lexém a pomocou pravidiel ich diferencuje od substantívnych významov: infinitív

[2] Na základe teórie P. Sgalla rozvíjanej jeho žiakmi sa v súčasnosti vo svetovej lingvistike všeobecne prijíma, že aktuálne členenie vety patrí medzi javy relevantné pre jej význam.

[3] V slovenčine by sa na prvý pohľad mohlo zdať, že infinitívne zakončenie na *-ť* vylučuje homonymiu s inými tvarmi, ale máme minimálne dve lexémy, ktoré sú ako slovesá veľmi frekventované, a to aj v infinitíve, a súčasne homonymné so substantívami v N a A tvare: *mať, stať (sa)* – substantívum *mať* vo význame matka, mater, substantívum *stať* vo význame článok, väčšia časť textu. Napr. lema *mať* sa v korpuse prim-4.0-public-all vyskytuje 2 414 658 x, z toho tvar *mať* 199 939 x. Značku S (substantívum) má automatizovane priradenú 1 522 týchto slov, ale ručným triedením zisťujeme, že iba približne 30 – 40 % z nich je naozaj správne označené substantívum *mať*. Automatizovaná anotácia založená iba na štatistike nemôže byť v takýchto prípadoch dostatočne spoľahlivá.

nemôže mať pred sebou predložku ani adjektívum, nemôže mať genitívnu valenciu a pod. Exaktne formulované pravidlá sa prevádzajú do podoby počítačových pravidiel a zlepšujú lingvistickú anotáciu textov v korpuse. Ak V. Petkevič v úvode svojho príspevku konštatoval, že „infinitiv je slovní tvar dost nevděčný" (s. 228), tak A. Rosen ukázal, že doplnok infinitívu a jeho pád je jav teoreticky oveľa zložitejší a náročný aj na správne použitie v praxi. Už vyhľadávanie zodpovedajúcich konštrukcií v korpuse (typu *Marie odnaučila Honzu přicházet* ***opilý/opilého***.; *Mobilita stylu umožnila Čapkovi být zároveň* ***obsažný*** *i* ***sdělný***.) je problematické vzhľadom na to, že pád pri závislom infinitíve sa neriadi všeobecnými pravidlami, ale môže sa líšiť aj podľa morfosyntaktických kategórií zúčastnených vetných členov. Okrem korpusových metód autor využíval aj introspekciu a elicitáciu a zamýšľal sa nad výhodami a nedostatkami každého postupu. Automatickým určením základných (morfo)syntaktických funkcií tvarov slovesa *byť* v češtine sa v príspevku *Pomocné sloveso být a automatická identifikace jeho hlavních funkcií* (s. 131 – 152) zaoberali Milena Hnátková a V. Petkevič. Autori si vytýčili za cieľ správne identifikovať sloveso *byť* a jeho tvary vo funkcii pomocného slovesa na tvorenie minulého času, opisného pasíva a opisného futúra nedokonavých slovies. Zameriavali sa na jednoduché vety, v ktorých nie sú čiarky ani iné viacvýznamové interpunkčné znamienka (vydeľujúce elipsu, parentézu, vnorenú vedľajšiu vetu a pod.), ktoré by komplikovali automatické určovanie funkcií tvarov slov. Každé z tvrdení o funkcii slovesa *byť* autori generalizovali pomocou vzorca spolu s vysvetlením a diskusiou k tvrdeniu. Predpokladajú, že sa im podarilo predstaviť komplexný výpočet všetkých možných variantov funkcií slovesa *byť*. V niektorých prípadoch je konečný výpočet problematický – najmä vo funkcii spolutvorby opisného futúra nedokonavých slovies vzhľadom na viacfunkčnosť futurálnych tvarov slovesa *byť* a infinitívu nedokonavých slovies. Výskum priniesol nové poznatky, ktoré boli po implementovaní do počítačových programov použité najmä pri tvorbe Frekvenčného slovníka češtiny (2004).

Špecifickým problémom – *Neshoda adjektiva s následujícím substantivem* (s. 153 – 179) – sa zaoberal Tomáš Jelínek. Na lepšie objasnenie problematiky autor podrobne zhrnul typy adjektív v češtine, predovšetkým však roztriedil výskyty spojení adjektívum a substantívum zo syntaktického hľadiska (priame a nepriame), výskyty týchto spojení ako súčasť nominálnej frázy, ale aj nezhody spôsobené rôznymi chybami. Cieľom príspevku bolo okrem podrobnej exemplifikácie chýb v anotácii bigramov adjektívum – substantívum prispieť aj k odstráneniu tohto typu chýb v automatickej morfologickej anotácii dát v ČNK.[4]

[4] Podobnú analýzu na materiáli Slovenského národného korpusu a na zlepšenie jeho anotácie realizovala K. Gajdošová (2010). Okrem chýb priamo v texte (preklepy, rozdelené slová) a chybného označenia jedného z členov bigramu vyčlenila aj skupinu lingvisticky zaujímavých a pre slovenčinu nie celkom typických prípadov rozdelenej syntagmy.

Do súboru syntaktických príspevkov môžeme zaradiť aj príspevok F. Čermáka *Lexikon nebo syntax? Nechce se mi a sestry její* (s. 68 – 94). Autor v ňom informuje o existencii osobitnej skupiny viet, ktoré sú problematické pre teoretickú syntax a ktorým sa bádatelia bežne nevenujú, hoci ich tvoria veľmi frekventované slovesá. Ide o typy viet s formálnou absenciou subjektu, s kombináciou reflexívneho *sa* a datívu a pod., napr. *Chce se mi spát.; Dělá se mi nanic.; Tobě se to řekne!* Autor tieto javy kategorizuje a kladie si otázku, či sú uvedené a im podobné konštrukcie syntakticky rozložiteľné alebo ide o už hotové skupiny slov, po ktorých používateľ siahne, ad hoc ich použije, príp. ich rozvinie či doplní. Prehľad názorov na uvedené konštrukcie dopĺňa autor porovnaním jednotlivých skupín s podobnými javmi (najmä s ergatívom) v iných jazykoch, ale aj v starej češtine. V závere F. Čermák konštatuje, že je nevyhnutné načrtnuté triedy syntakticky problematických javov analyzovať na báze korpusu a uvádzať ich v gramatikách a slovníkoch explicitne komplexným výpočtom.

Na pomedzí lexikónu a syntaxe sa nachádzajú aj príspevky o adjektívach (Marie Kopřivová: *Kolokační profil nejčastějších adjektiv v korpusech ČNK*, s. 180 – 204[5]) a časticiach (Renata Blatná: *Částice v kontextovém okolí víceslovných předložek*, s. 36 – 52; Martin Stluka: *Příklonné částice v textech počátků české prózy*, s. 314 – 329). R. Blatná predstavila jednak formálnu analýzu časticovo-predložkových kombinácií, jednak sa pokúsila o sémantickú kategorizáciu častíc v závislosti od pozície v kontextovom okolí viacslovných predložiek (na základe klasifikácie F. Čermáka). Pri viacslovných predložkách sa podľa jej zistení najčastejšie nachádzajú emocionálno-intenzifikačné častice *především, zejména, i*, ktoré stoja najmä pred predložkami s významom determinácie (*na základě, v případě*), času (*v době*) a kauzality (*v důsledku*). M. Stluka vo svojom diachrónnom bádaní tentoraz dokazuje, že príklonné častice *t*-ového a *ž*-ového charakteru (*jakžtotě, jěz že, kamžť*) patria na konci 14. st. medzi značne využívané jazykové prostriedky, pričom ich funkcia nie je obmedzená na zosilňovanie významu lexém, ale vyskytujú sa ešte v deiktickom význame, ba bolo by možné uvažovať aj o vetnočlenskej platnosti tohto typu častíc.

Špecifický diachrónny výskum na materiáli korpusu vytvoreného z výberu z českých textov od polovice 13. st. (dohromady 2 mil. slov) predstavil Karel Kučera: *Kvantitativní charakteristika průběhu a uplatnění změn ý>ej, ú>ou a aj>ej v češtině od 14. století do současnosti* (s. 210 – 225). Obmedzený rozsah materiálu a analýza iba troch hláskových zmien mu síce neumožňujú robiť všeobecné závery, no predsa sa autorovi ukazuje, že priebeh a uplatnenie sledovaných hláskových zmien sa neriadi jedinou hierarchiou troch základných pozícií (začiatočná, stredová, koncová), ale jednotlivé zmeny sa na nich realizujú nerovnakým spôsobom.

[5] Postupy a výsledky M. Kopřivovej budú podrobnejšie opísané v recenzii jej monografie o adjektívnej valencii, ktorá vyšla ako 4. zväzok edície Štúdie z korpusovej lingvistiky.

Kým v oblasti histórie jazyka je vždy nedostatok jazykového materiálu, pri analýze súčasných písaných textov zhromaždených v korpuse už zvyčajne nastáva problém s veľkým rozsahom vyhľadaných dokladov, ako sme spomínali aj vyššie. Michal Šulc vo svojom príspevku *Frekvence jevu v korpusu a dva typy jejího referenčního rámce* (s. 330 – 346) ponúka námety, ako využiť funkciu absolútnej frekvencie pri práci na báze korpusu. Na príklade výskumu slovotvorného formantu *-dlo* s významom miesta vysvetľuje v niekoľkých krokoch nadväznosť jednotlivých negatívnych filtrov tak, aby získal čo najoptimálnejší materiál bez prítomnosti nerelevantných výskytov (nečeská majuskula, interpunkcia a pod.). Autor prezentuje jednu z možností spracovania skúmaného javu a vyzdvihuje existenciu viacerých postupov práce s materiálom a jeho filtrácie. V závere vyčísľuje komunikačnú zaťaženosť zvoleného formantu a porovnáva ju s inými konkurenčnými formantmi (*-ovka, -ovna*). Dostatok materiálu nechýbal ani v príspevku *Určení jazykové základovosti barev v Českém národním korpusu* (s. 285 – 313). Věra Schmiedtová a Barbara Schmiedtová v ňom dopĺňajú a rozširujú starší výskum s podobným zameraním, pričom sa sústreďujú na zvyšné farby zo skupiny základových farieb (zelená, modrá, žltá, šedá/šedivá, ružová, hnedá, fialová, oranžová). Pomenovania týchto farieb skúmajú zo sémantického a frekvenčného hľadiska, všímajú si ich zapojenie do jazykového systému ako napr. tvorenie derivátov, frazém, termínov, ale aj spracovanie výkladov jednotlivých farieb v štyroch základných českých slovníkoch. Príspevok je bohatý na grafy a tabuľky s vyčíslením výskytov jednotlivých farieb a ich derivátov, nimi tvorených frazém a termínov, ich sémantickej zaťaženosti a pod. Vzhľadom na vysokú frekvenciu a spájateľnosť vyčleňujú autorky len malý priestor aj odtieňom týchto základových farieb.

Osobitnú tému predstavujú v zborníku príspevky o hovorených korpusoch. F. Čermák sa v štúdii *Mluvené korpusy* (s. 53 – 67) zameriava na ich typológiu (špecializovaný korpus, všeobecný korpus) a venuje sa aj aspektom a faktorom, ktoré výstavbu hovorených korpusov ovplyvňujú. Ide o jazykové (čas a priestor, vzťah autora a adresáta, funkcia a cieľ komunikácie, štruktúra prehovoru) i nejazykové aspekty. Časť príspevku venoval autor problému získavania dát do hovoreného korpusu z médií a stratégii terénneho zberu. Prehľadovo je spracovaná pasáž o existencii a vývoji niektorých hovorených korpusov vo svete a v ČR. F. Čermák podrobne analyzuje aj kritériá stratifikácie hovoreného korpusu, osvetľuje metadáta – informácie o respondentovi a nahrávke, ktoré považuje v hovorenom korpuse za nevyhnutné. Rozsiahla príloha k článku pozostáva z ukážky prepisu zvukovej nahrávky z Pražského hovoreného korpusu, z komentovaného frekvenčného slovníka a zoznamu kolokácií z uvedenej ukážky. François Esvan v článku *Srovnávací rozbor mluvených korpusů (PMK a BMK): metodologické problémy a první výsledky* (s. 95 – 117) podrobne vysvetľuje štruktúru Pražského mluveného korpusu a Brnenského mluveného korpusu z hľadiska metadát. Autor matematicky analyzuje koeficient korekcie

pre jednotlivé korpusy pri práci s dátami na porovnávacie výskumné ciele s použitím štatistických nástrojov. Metodologické úvahy overuje na príklade výskumu koncovky *-ej* v PMK a BMK. V závere upozorňuje na skreslené výsledky porovnávacieho výskumu v prípade nerešpektovania korekcie, ale aj na nevyhnutnosť ďalšieho materiálového dopĺňania databáz na porovnávacie výskumy. Výsledok modelového výskumu F. Esvana nabáda tvorcov korpusov na doplnenie nových kategórií do metadát o respondentoch. Martina Waclawičová príspevkom *Mluvené korpusy v ČNK: několik poznámek k mluveným projevům a polyfunkčním výrazům* (s. 347 – 358) dopĺňa predchádzajúce informácie o pohľad do českých hovorených subkorpusov a o zhrnutie základných vlastností spontánnej komunikácie ako situačná ukotvenosť, výskyt kontaktových, nadväzovacích, výplnkových výrazov, nepripravenosť, opakovanie, opravy a nedokončené vety. Autorka sa sčasti zmieňuje aj o koncepčnej príprave jednotlivých českých hovorených korpusov, najmä Českého mluveného korpusu, o spolupráci s vysokými školami, ale aj o problémoch s vyváženosťou hovorených subkorpusov. *Tvary minulého příčestí v ČNK: táh, táhl, či táhnul?* (s. 19 – 35) skúmal Neil Bermel na písaných aj hovorených korpusoch češtiny. Jeho cieľom bolo zhodnotiť súčasný stav variantnosti v tomto type, zistiť jeho funkčné rozlíšenie a konfrontovať gramatické výklady so sondami do materiálu ČNK. Ukázalo sa, že reálne využitie rôznych variantov, ako dokladujú korpusové údaje, vcelku zodpovedá gramatickému opisu, ktorý by však bolo možné doplniť informáciami o využití a funkciách aj okrajovejších tvarov.

V zborníku ***Korpusová lingvistika: Stav a modelové přístupy*** sú zhrnuté metodologické i analytické príspevky, ktoré odrážajú stav korpusovolingvistického výskumu v Českej republike v r. 2005 – 2006. Vývoj vedy a výskumu je v súčasnosti oveľa dynamickejší než bol ešte pred 15 – 20 rokmi, čo osobitne platí o nových, akceleračne sa rozvíjajúcich odboroch, akým je aj korpusová lingvistika. Viacerí z prezentovaných autorov medzitým rozvinuli svoje analýzy a pozorovania do syntetických monografických prác, spolupodieľali sa na príprave 1. zväzku Mluvnice současné češtiny (2010) založenej na reálnych dátach z písaných a hovorených korpusov. Ústav Českého národného korpusu vydal okrem toho ďalšie frekvenčné slovníky (hovorenej češtiny, autorské – Karla Čapka, Bohumila Hrabala, ako aj slovník jazyka z obdobia komunistickej totality), rozšírili a skvalitnili sa korpusové zdroje, k dispozícii sú nové anotačné procedúry a vyhľadávacie nástroje. Edícia Studie z korpusové lingvistiky a jej jednotlivé zväzky však môžu byť aj s odstupom času metodologicky inšpiratívne a výsledkami realizovaných výskumov môžu poslúžiť pri komparatívnom štúdiu slovenčiny a češtiny, ako aj iných jazykov. Využiť sa pritom dá jednak rozsiahly dokladový materiál uvedený v zborníku, jednak jedno z nesporných pozitív každého korpusu: možnosť opakovanej analýzy toho istého materiálu a verifikácie výsledkov.

Literatúra

CVRČEK, Václav a kol.: Mluvnice současné češtiny. I. Praha: Nakladatelství Karolinum 2010. 354 s.
Český národní korpus – úvod a příručka uživatele. Ed. J. Kocek – M. Kopřivová – K. Kučera. Praha: Ústav Českého národního korpusu FF UK 2000. 159 s.
Frekvenční slovník češtiny. Ed.: F. Čermák – M. Křen. Praha: Nakladatelství Lidové noviny 2004. 596 s.
GAJDOŠOVÁ, Katarína: Lingvistická analýza automatizovanej anotácie datívu adjektíva v heterogénnych menných skupinách v Slovenskom národnom korpuse. In: Slovo – Tvorba – Dynamickosť. Ed.: M. Šimková. Bratislava: Veda 2010, s. 309 – 330.
Gramatika a korpus / Grammar & Corpora 2005. Ed. F. Štícha – J. Šimandl. Praha: Ústav pro jazyk český AV ČR 2007. 304 s.
Korpus jako zdroj dat o češtině. Ed. P. Karlík. Brno: Masarykova univerzita 2004. 208 s.
Možnosti a meze české gramatiky. Ed. F. Štícha. Praha: Academia 2006. 304 s.
Studie z korpusové lingvistiky. Ed. F. Čermák – J. Klímová – V. Petkevič. Acta Universitatis Carolinae. Philologica 1997, č. 3 – 4. Praha: Karolinum 2000. 531 s.
ŠIMKOVÁ, Mária: Výberový slovník termínov z počítačovej a korpusovej lingvistiky. In: Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Ed. M. Sokolová – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity 2006, s. 187 – 193. Dostupný aj na www: http://korpus.juls.savba.sk/publications/block1/2006-simkova-vyberovy%20slovnik%20terminov/2006-simkova-vyberovy%20slovnik%20terminov.pdf
ŠULC, Michal: Korpusová lingvistika. První vstup. Praha: Karolinum 1999. 94 s.

KORPUSOVÝ POHĽAD NA SPÁJATEĽNOSŤ SLOV

(Kolokace. Studie z korpusové lingvistiky. Svazek 2. Ed. F. Čermák, M. Šulc. Praha: Nakladatelství Lidové noviny – Ústav Českého národního korpusu 2006. 452 s. ISBN 80-7106-863-2)

Daniela Majchráková
Slovenský národný korpus Jazykovedného ústavu Ľudovíta Štúra SAV
Panská 26, 813 64 Bratislava
e-mail: danam@korpus.sk

V roku 2006 vyšiel druhý zväzok Studií z korpusové lingvistiky s názvom Kolokace. Táto zbierka 16 štúdií je prvým českým zborníkom venovaným danej tematike, hoci čiastkové výstupy sa objavili už skôr.[1]

V úvodnom príspevku *Kolokace v lingvistice* (s. 9 – 16) František Čermák čitateľa oboznamuje so základnými poznatkami týkajúcimi sa kolokácií, čím mu poskytuje potrebnú teoretickú bázu pre jednotlivé štúdie zamerané na čiastkové otázky. Kolokácie autor považuje za kľúč k poznávaniu spôsobov, pravidiel a hraníc správa-

[1] Spomeňme napríklad štúdiu F. Čermáka Syntagmatika slovníku: typy lexikálních kombinací. In: Čeština – univerzália a specifika 3. Eds. Z. Hladká, P. Karlík. Brno: Vydavatelství Masarykovy univerzity 2001, s. 223 – 232.