Hladká, Zdeňka

[Šulc, Michal. Korpusová lingvistika: první vstup]

Sborník prací Filozofické fakulty brněnské univerzity. A, Řada jazykovědná. 2000, vol. 49, iss. A48, pp. 194-196

ISBN 80-210-2350-3
ISSN 0231-7567

k osvojení, doplnění a rozvinutí představených názorů a k novému promýšlení tradičních pojmů. V okruhu teoretických sociolingvistických prací se tak monografie stává významným přínosem. Vydání mimo slovanské země však bohužel osloví více zahraniční slavisty než ty, kdo živé jazykově dění studují v bezprostředním kontaktu s realitou.

Marie Krčmová

Michal Šulc: **Korpusová lingvistika. První vstup**. Univerzita Karlova — nakladatelství Karolinum, Praha 1999, 94 str. (ISBN 80–7184–847–6)

Útlá knížka Michala Šulce, pracovníka Ústavu Českého národního korpusu v Praze, je koncipována jako učební text pro posluchače Filozofické fakulty Univerzity Karlovy. Přináší přehledně, srozumitelně a zajímavě zpracované poučení o relativně mladé, avšak rychle se rozvíjející a stále významnější disciplíně jazykovědného bádání — korpusové lingvistice. V předmluvě autor říká, že jeho cílem nebylo podat soustavný výklad celé problematiky ani zasvětit čtenáře do detailů práce odborníků vytvářejících jazykové korpusy. Chtěl pouze „pomoci objevit nové možnosti práce s jazykem, upozornit na nový typ informační databanky a vzbudit zájem o nový obor" (s. 7). Domníváme se, že právě toto pojetí bylo zvoleno šťastně. V českém prostředí se sice v posledních letech objevila řada studií a článků o korpusové lingvistice (např. Čermák, F.: *Jazykový korpus. Prostředek a zdroj poznání.* SaS, 56, 1995, s. 119–140; Čermák, F.: *Komputační lexikografie.* In: Čermák, F. — Blatná, R. /eds./, Manuál lexikografie. H&H, Jinočany 1995, s. 50–71; Čermák, F. — Králík, J. — Kučera, K.: *Recepce současné češtiny a reprezentativnost korpusu.* SaS, 58, 1997, s. 117–124; Blatná, R.: *Textové korpusy slovanských jazyků.* In: Slavica Pragensia ad tempora nostra, Praha 1998, s. 190–194), souhrnné poučení přístupné i laickým zájemcům však dosud chybělo.

Práce M. Šulce je rozdělena do sedmi kapitol. První tři vysvětlují obecné pojmy týkající se korpusové lingvistiky, informují o historii této mladé disciplíny i o jejím současném postavení ve světě. Druhá polovina knihy ve čtyřech kapitolách charakterizuje Český národní korpus, jeho tvorbu, specifika i možnosti využití.

Východiskem první kapitoly nazvané *Korpusová lingvistika* je autorův pokus o co nejadekvátnější definici pojmu jazykový korpus (s. 11): „Korpus je rozsáhlý soubor elektronických textů, cíleně shromážděný jako referenční zdroj pro vědecké studium jazyka a pro zpracování užitných jazykových nástrojů, který je v jednotném formátu, je lingvisticky označkovaný a který lze z hlediska skladby považovat za jistým způsobem vyvážený." Dále jsou popsány jednotlivé typy korpusů (k. jednoho jazyka X k. paralelní; k. synchronní X k. diachronní; k. všeobecné X k. specificky zaměřené; k. uzavřené X k. otevřené, tzv. monitorovací; k. psaného jazyka X k. mluveného jazyka ad.), detailně je probírána dosud ne zcela dořešená otázka reprezentativnosti korpusů, tj. volby kvantitativních a kvalitativních kritérií pro výběr ukládaných textů, apod. Čtenáři jsou též seznámeni se softwarovými nástroji, s jejichž pomocí je možno s korpusy pracovat.

V kapitole nazvané *Historie korpusové lingvistiky* je podán přehled o rychlém vývoji disciplíny od skromných počátků v 60. letech 20. století (stručně jsou zaznamenány i kořeny starší) do současného rozvinutého stavu. Popis je zaměřen na anglicky mluvící prostředí, v němž korpusy vznikly a v němž jsou i dnes nejrozšířenější. Dozvídáme se např., že prvním elektronickým korpusem vytvořeným primárně pro lingvistické účely byl *Brown University Standard Corpus of Pre-*

sent-Day Edited American English (nazývaný též *Brown Corpus*), sestavený v letech 1961–1964 H. Kučerou a N. Francisem. Rozsahem se první korpus zdaleka nerovnal projektům současným (cílem bylo shromáždit pouze milion slov psaného textu), jeho pečlivě promyšlená struktura však nadlouho představovala standard pro podobné práce. Po řadě menších korpusů ze 70. a 80. let otevřel éru velkých korpusů projekt *Collins Birmingham University International Language Database* (označovaný *COBUILD Corpus*), jehož počátky jsou datovány rokem 1980. V roce 1990 byl překoncipován v první monitorovací korpus na světě (*Bank of English*) s předpokládaným růstem na několik set milionů slovních tvarů (na konci devadesátých let obsahoval téměř 400 milionů slov, práce na tomto projektu však byla pozastavena). Korpusem, který se svým reprezentativním složením, propracovaností struktury i mnohovrstevnou klasifikací materiálu stal novým standardem pro současnou korpusovou tvorbu, je *British National Corpus*, sestavený v letech 1991–1995. Obsahuje 100 milionů slov, z nichž 10 % tvoří mluvená angličtina (největší subkorpus mluveného jazyka na světě).

Třetí kapitola přibližuje některé specializované korpusové projekty, autor se snaží zachytit hlavní současné trendy disciplíny, upozorňuje na zajímavé iniciativy apod. Ve stručnosti se též zmiňuje o rozvoji korpusové lingvistiky mimo rámec anglofonních zemí a podává přehled zatím relativně malého zapojení slovanských zemí do této oblasti. Práce českých korpusových lingvistů se v předloženém srovnání jeví jako nejpokročilejší v slovanském světě.

Výklad o korpusové lingvistice v domácím českém prostředí je otevřen čtvrtou kapitolou nesoucí zastřešující název *Český národní korpus*. Její podtitul *Ústav Českého národního korpusu* napovídá, že se v ní hovoří o vzniku prvního českého korpusového pracoviště. Dozvídáme se, že už v roce 1988 byla pod záštitou Kybernetické společnosti ustavena „Iniciativní skupina pro přípravu počítačových korpusů textů a slovníků", v únoru 1991 se několik lingvistů a matematiků z Prahy a z Brna rozhodlo společnými silami vybudovat „Počítačový fond češtiny" a v rámci aktivit tohoto sdružení uzrála myšlenka vytvořit projekt celonárodního charakteru — Český národní korpus (ČNK). V září 1994 začal existovat Ústav Českého národního korpusu (ÚČNK), který od října 1996 získal vlastní pracoviště na Filozofické fakultě UK v Praze. Jeho vedoucím se stal F. Čermák. V rámci komplexního grantu z roku 1996 si ÚČNK klade za cíl „vytvoření korpusu češtiny (tj. elektronicky uloženého, zpracovávaného a přístupného souboru jazykových dat ve standardizovaném formátu), který by se stal univerzálním referenčním zdrojem pro vědecké studium češtiny i plně reprezentativní základnou pro vytvoření velkého slovníku českého jazyka a pro zpracování dalších jazykových příruček a užitných jazykových nástrojů" (s. 46n.).

Pátá kapitola nazvaná *Synchronní korpus* je věnována tvorbě, charakteru a možnostem využití nejdůležitější složky ČNK, tj. korpusu současných psaných textů (publikovaných převážně v období od r. 1990). V konečné podobě by měl obsahovat sto milionů slovních tvarů, měl by být co nejreprezentativnější, vyvážený, kvalitně označkovaný a jednoduše přístupný. Výklad M. Šulce podává poměrně detailní informace o jednotlivých fázích tvorby korpusu: od získávání textů přes jejich evidenci, archivaci, různé stupně konverze a čištění až k vnější lingvistické anotaci a gramatickému značkování. Ukazuje též, co a jakým způsobem je možno v synchronním korpusu ČNK vyhledat už dnes.

V šesté kapitole je pod názvem *Mluvený korpus* uvedena stručná informace o zpracování mluveného jazyka pro ČNK. Detailněji je přiblížen subkorpus mluvené češtiny z pražského prostředí (obsahující přepis 300 nahrávek, tj. zhruba 700 000 slovních tvarů), který se má stát mj. základnou pro připravovaný frekvenční slovník. O paralelně vznikajícím subkorpusu mluveného jazyka z brněnského prostředí a o tvorbě nástrojů pro jeho automatické morfologické značkování měl autor knihy k dispozici pouze povšechné a zčásti nepřesné informace, proto se mu věnuje jen okrajově. (V dohledné době tuto problematiku přiblíží např. článek absolventky Filozofické

fakulty Masarykovy univerzity D. Hlaváčkové *Korpus mluvené češtiny z brněnského prostředí*, který je připraven pro časopis Slovo a slovesnost.)

Poslední kapitola *Diachronní korpus* je věnována tvorbě samostatné složky ČNK, která by měla v budoucnu obsáhnout reprezentativní výběr textů od druhé poloviny 13. století po současnost, a to v hustotě zhruba 5 sond na desetiletí (s výjimkou nejstarších vývojových fází češtiny). Vzhledem k obtížnosti elektronického přepisu historických textů a též v důsledku prozatímní neuzpůsobenosti softwarových nástrojů k jejich dalšímu zpracování je tvorba diachronního korpusu teprve v začátcích. K dispozici je zatím pouze malý subkorpus o 16 sondách.

Práci M. Šulce doplňuje výběrový slovníček použitých lingvistických pojmů, slovníček pojmů z korpusové lingvistiky a velmi užitečný přehled nejdůležitějších institucí, projektů a softwarových nástrojů, které se týkají tvorby a zpracování korpusů. Nechybí ani seznam základní literatury oboru.

Anotovaná knížka dobře plní svůj účel: podává základní informace o korpusové lingvistice, aniž by příliš specializovaným výkladem odrazovala dosud nepoučené čtenáře, naopak povzbuzuje jejich zájem a poskytuje jim orientaci pro případné další vzdělávání v této oblasti. Pozitivně lze hodnotit také fakt, že se autor snaží vidět možnosti korpusové lingvistiky objektivně, tj. ukazuje její velký přínos pro popis jazyka, ale též hranice využití.

Zdeňka Hladká

Michael Betsch: **Diskontinuität und Tradition im System der tschechischen Anredepronomina (1700–1850).** Verlag Otto Sagner, München 2000, 198 s. Slavistische Beiträge 389. (ISBN 3-87690-754-3)

Oslovování, tedy způsob, jak se mluvčí obrací k adresátovi, má význam jak směřující mimo jazyk (odkazující k sociální strukturaci jazykové komunity), tak i směřující do jazyka (to je dobře vidět na příkladu japonštiny, kde místo kategorie osoby je systém různých pojmenování mluvčího, adresáta a ostatních). Soustavné zkoumání vývoje oslovování v češtině dosud chybělo. Přináší je až recenzovaná monografie, která byla přijata jako disertace na univerzitě Tübingen (kniha má vročení 2000, ale ve skutečnosti vyšla už v listopadu 1999).

Jak již naznačuje titul knihy, centrální otázkou je role národního obrození ve vývoji oslovování v češtině. Svoje závěry autor staví na důkladném rozboru empirického materiálu (kap. 3). Jeho pramennou základnu je možno rozdělit zhruba na tři okruhy. Jednak jsou to prameny normativní (tedy svou povahou — v bernheimovské terminologii — záměrné); sem patří dobové gramatiky od Rosy až po Burianovu Böhmische Sprachlehre, jakož i vzorové dialogy připojované k některým gramatikám i vydávané samostatně. Dále byly excerpovány prameny odrážející reálnou komunikaci: dobová korespondence (ta může být nezáměrná i záměrná, jako je tomu v případě dopisů představitelů národního obrození) a některé další prameny, jako protokoly soudních výslechů. Třetí okruh pak sestává z dobové umělecké literatury (zejm. divadelních her). Všechny tyto prameny byly podrobeny velice subtilní analýze, vědomé si všech jejich jednotlivých zvláštností vyplývajících z jejich různé povahy. Výsledkem je rekonstrukce vývoje oslovování v daném období (kap. 4). Tu lze poněkud zjednodušeně shrnout asi následovně:

Kolem roku 1700 existuje systém s třemi způsoby oslovovaní. Nejprestižnější je nepřímé oslovení se substantivem *pán* (to je doloženo už v Blahoslavově gramatice z r. 1571), prostřední pozici zaujímá vykání (ovšem ještě s plurálovou kongruencí), nejníže je tykání. Kolem poloviny