FRANTIŠEK ČERMÁK: *KORPUS A KORPUSOVÁ LINGVISTIKA*

Praha: Nakladatelství Karolinum, 2017, 268 stran
ISBN 978-80-246-3710-5

Autor je jedním z hlavních iniciátorů korpusové lingvistiky v České republice. Snahy o vytvoření českého textového korpusu spadají do začátku 90. let v souvislosti s ustanovením občanského sdružení Počítačový fond češtiny (mezi jehož členy patřila i E. Hajičová, K. Pala a další) v r. 1991. Po té, co se myšlenka na korpus nesetkala s pochopením na půdě Ústavu pro jazyk český, obrátil se František Čermák na FF UK, kde mu v r. 1994 bylo umožněno založit tehdy skromný Ústav českého národního korpusu. Ústav se postupně rozrostl a díky tomu, že se Čermák dokázal obklopit schopnými mladými lidmi a nadchnout je pro obor, stalo se z něj špičkové pracoviště, srovnatelné s obdobnými zahraničními centry. V současnosti představuje ÚČNK jedno z nejúspěšnějších a nejprogresivnějších pracovišť na FF UK. Za dobu existence ústavu zde vznikla série českých synchronních korpusů SYN (2000 až 2015, syn v3-6), linie českých mluvených korpusů (PMK, řada ORAL a nedávno zveřejněné korpusy ORTOFON a DIALEKT) a paralelní korpusy v rámci projektu InterCorp.

Cílem publikace není podat zevrubné informace o korpusové lingvistice a jejím současném stavu. Vzhledem k obrovské šíři tohoto odvětví lingvistiky, jeho současné teoreticko-metodologické úrovni a překotnému vývoji by to byl pro jednotlivce úkol téměř nadlidský. Svědčí o tom fakt, že podobných základních úvodů např. v angličtině, s níž je korpusová lingvistika neodmyslitelně spjata, vyšlo za posledních dvacet let relativně velmi málo (mezi nejznámější a ty nejnovější patří McEnery — Wilson, 1996/2001; Biber — Conrad — Reppen, 1998; Cheng, 2012; McEnery — Hardie, 2012; Weisser, 2016; Desagulier, 2017).

Autor knihu pojal jako oborovou příručku a sborník zároveň a rozdělil ji do tří částí, které doplňuje glosář základních pojmů a termínů. První část, *I. Úvod: teorie*, obsahuje šest oddílů a zabírá polovinu knihy. První oddíl, *Úvod: korpus a korpusová lingvistika*, shrnuje záměr publikace: podat šířeji pojatý nástin postavení a funkce korpusové lingvistiky v moderní lingvistice a jejího přínosu pro ni, přehled základních pojmů a teoretických a praktických otázek až po konkrétní studie a ukázky. Text je výsledkem autorovy produkce za dlouhou řadu let, některé pasáže čerpají z autorových encyklopedických textů a poslední třetí část je založena na jeho empirických studiích.

Druhý oddíl, *Data a informace*, se zabývá tím, z čeho korpusová lingvistika vychází, tzn. informačními jednotkami uloženými v elektronických textech. Jednotlivé pododdíly popisují povahu a zdroje elektronických dat, charakter, vlastnosti a zpracování textů, v nichž jsou elektronická data uložena, včetně recepce a produkce těchto souborů textů, tj. korpusů. Další oddíly referují o jejich sestavování a o přípravě korpusových dat v nich obsažených za účelem jejich efektivního vytěžování pomocí dotazů a dotazovacího jazyka. Třetí oddíl, *Korpus a korpusy*, jak naznačuje název, je zaměřen na vytváření korpusů, jejich typologii, charakteristiky a hledání v nich. Popisuje základní způsoby extrakce dat, konkordance, a typicky vyhledávané jednotky, kolokace, n-gramy a koligace, a jejich analýzu. Ukazuje, jak se z korpusových dat, jejichž prohledávání je z povahy věci založeno na formě, dospívá ke zjištěním o jejich významu a funkci. Další pododdíly probírají specifické typy korpusů: synchronní,

diachronní, mluvené, vícejazyčné paralelní, resp. překladové (významný projekt InterCorp párující přes 30 jazyků vzniká v ÚČNK od r. 2005) a webové (tj. soubory textů stažené z internetu o rozsahu miliard slov, ovšem s řadou nevýhod). Závěrem podává přehled o hlavních světových korpusech v ostatních jazycích, které rozděluje do tří generací (první dvě se až na pár výjimek týkají angličtiny; kupodivu v žádné z těchto tří generací není zmíněn ČNK).

Čtvrtý oddíl první části, *Korpusová lingvistika*, představuje náplň tohoto odvětví lingvistiky, jeho metodologii, základní pojmy (včetně klíčových slov, poměr type-token ad.), operace a statistické postupy. Následující pátý oddíl, *Korpus ve výzkumu a aplikaci*, vedle otázek spojených s výzkumem obecně a s výzkumem lingvistickým zběžně probírá i nové možnosti zkoumání variability jazyka, zvláště pak mluveného, prostřednictvím korpusů. Z oblastí, na které je korpusový výzkum aplikován, jsou zahrnuty lexikologie a lexikografie, ale též gramatika, morfologie a syntax, a to z hlediska zkoumání jednoho i více jazyků (kontrastivní výzkum), přičemž oba přístupy nacházejí významné uplatnění v pedagogické praxi. V šestém, závěrečném oddílu se autor stručně věnuje perspektivě korpusové lingvistiky. Kromě zdokonalování korpusů a metodologie vidí hlavně těžiště dalšího výzkumu v zaměření se na syntagmatiku, její jevy a zákonitosti, tedy oblast, která byla v minulosti nejvíce omezena nedostatkem kvantitativních dat. Vedle pokročilého zkoumání kombinatoriky lexémů jedno- i víceslovných počítá s přechodem ke kombinatorice větších útvarů, prefabrikovaných součástí různých typů textů a promluv, a k hledání hranic mezi pravidelnými (pravidlům podléhajícími) oblastmi jazyka a jazykem nepravidelným.

Druhá část publikace, *Sondy, studie, analýzy*, předkládá autorovy konkrétní již vydané studie, a to obecnější i materiálové. První z nich se vrací k typologii korpusů a sleduje jejich vývoj v průběhu několika desetiletí až do přítomnosti a načrtává některé pravděpodobné možnosti vývoje (integrace popisu lexikonu a gramatiky) a specifické úkoly (dokonalejší popis jazykové reality, dokonalejší pokrytí paralelních jazykových dat a řešení otázky potenciality jazykových dat v návaznosti na zkoumání metaforičnosti jazyka). Další popisuje patnáct desiderat a postřehů týkajících se praxe korpusového výzkumu (např. kvality a kvantity dat, nutnosti přímých dat, reprezentativnosti korpusu, problémů s tagováním a lemmatizací atd.). Následující studie se týká problematiky mluvených korpusů, jejich sestavování, parametrů a kritérií pro zařazování mluvených projevů (demografických, situačních apod.). Poslední z úvodních obecnějších statí je úvaha nad paralelním korpusem InterCorp, z něhož některé z následujících empirických studií vycházejí a jsou determinovány jeho charakterem a možnostmi. Materiálové studie se týkají lexikální kolokability sloves a adverbií, abstraktních substantiv i konkrétního slova (*konference*) a jeho vzrůstající polysémie. Zahrnuty jsou i studie metod vyhledávání idiomů v textových korpusech a mapování interjekcí v češtině prostřednictvím korpusových dat. Čtyři z těchto devíti sond v druhé části (B, C, E a H) jsou poněkud neorganicky vzhledem k ostatnímu textu ponechány v angličtině.

Poslední třetí část nazvaná *Přílohy* obsahuje čtyři krátké ukázky korpusových výstupů, často opět převzatých z jiných publikací. Jde o ukázku konkordance, hesel z Frekvenčního slovníku češtiny, tabulek ze Statistik češtiny a konečně výsledky korpusové analýzy konkrétního románu.

Nesporným kladem Čermákovy publikace je fakt, že je první svého druhu pro češtinu a dost možná dlouho zůstane osamoceným dokumentem o vývoji české korpusové lingvistiky v devadesátých letech a na začátku nového tisíciletí. Je ale také pravda, že publikaci lze leccos vytknout. Text je velkou měrou sestaven ze starších textů, které bylo možno s odstupem let od jejich vzniku pro tyto účely důkladněji adaptovat a revidovat (a v některých případech přeložit do češtiny). Mnoho věcí od doby, kdy většina použitých textů vznikala — tj. v první dekádě tohoto tisíciletí — prošla někdy menšími, někdy výraznými proměnami a některé priority a názory se posunuly. Např. při popisu ČNK na s. 69 se přímo píše, že jde o stav ke konci roku 2013, což je téměř čtyři roky před vydáním publikace. V důsledku kolážového skládání textu se některé popisy v různých částech knihy opakují a překrývají (např. popis textových formátů na s. 14 a 29), u převzatých pasáží není vždy uveden zdroj. Výhrady je možné mít také k výkladům některých pojmů, např. u ARF (average reduced frequency; s. 110) se uvádí, že „frekvenci koriguje tím, že ji přepočítává s ohledem na počet různých zdrojů, resp. textů". Asi přesnější by bylo říci, že rozděluje korpus do stejně velkých úseků podle frekvence daného slova a zohledňuje všechna možná sestavení (pořadí textů v) korpusu. Drobných výtek a redakčních přehlédnutí by bylo možno jmenovat více (neúplné citace v textu ap.).

Nicméně není od věci připomenout 5. bod Slovníkářova desatera (Čermák, 1995, s. 113), které platí i v tomto případě: Žádný slovník není bez chyb, protože slovník je umění možného. Přes všechny možné výhrady je dobré, že český čtenář má nyní k dispozici oborovou příručku o korpusové lingvistice, která přihlíží ke specifiku českých korpusů a české korpusové lingvistiky. Je na mladších korpusových specialistech, aby časem přišli s příručkou nové generace a přidali nové kapitoly.

LITERATURA

Biber, D. — Conrad, S. — Reppen, R. (1998): *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Cheng, W (2012): *Exploring Corpus Linguistics: Language in Action*. London and New York: Routledge.

Čermák, F. (1995): Paradigmatika a syntagmatika slovníku: možnosti a výhledy. In: F. Čermák — R. Blatná (eds), *Manuál lexikografie*. Praha: H+H, s. 90–11.

Desagulier, G. (2017): *Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics*. Springer International Publishing AG.

McEnery, T. — Hardie, A. (2012): *Corpus Linguistics. Method, Theory and Practice.* Cambridge: Cambridge University Press.

McEnery, T. — Wilson, A. (1998, 2nd ed. 2001): *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Weisser, M. (2016): *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis.* Malden: Wiley Blackwell.

Aleš Klégr | Ústav anglického jazyka a didaktiky, Filozofická fakulta Univerzity Karlovy | nám. Jana Palacha 2, 116 38 Praha 1
ORCID ID: 0000-0001-7760-6631
aleš.klegr@ff.cuni.cz