ČASOPIS PRO MODERNÍ FILOLOGII 103, 2021, Č. 2, S. 287–305

ANATOL STEFANOWITSCH: *CORPUS LINGUISTICS. A GUIDE TO THE METHODOLOGY*

Language Science Press, 2020, 490 stran
ISBN: 978-3-96110-224-2

OPEN ACCESS

Jazykové korpusy a s nimi neoddělitelně spjatá korpusová lingvistika už nějakou dobu patří mezi základní pilíře jazykovědného bádání, které si klade za cíl popisovat reálný jazyk a odhalovat jeho rysy a zákonitosti. V průběhu let byla o korpusovou lingvistiku a její status svedena nejedna argumentační bitva, fundamentálními námitkami generativistů počínaje a dílčími výtkami k reprezentativnosti korpusů konče. Není proto divu, že vznikla celá řada publikací, které se snažily korpusovou lingvistiku souhrnně představit a charakterizovat i s celou její historií a teorií: z posledních let zmiňme například Hardie — McEnery (2011) nebo Biber — Reppen (2015). Kniha německého lingvisty Anatola Stefanowitsche *Corpus linguistics. A Guide to the Methodology*, která vyšla v loňském roce, je však v mnoha ohledech ojedinělá.

Hned na první pohled zaujme a potěší, že publikace vyšla v režimu open access, tedy zdarma dostupná pro každého ke stažení, a jak sám autor nabádá, i k případnému doplnění či přizpůsobení potřebám každého zájemce. Už jen to, že autor nepovažuje svou knihu za neměnné a kanonické dílo, a naopak je otevřený případným doplněním ze strany svých čtenářů, je v rámci odborné literatury velmi osvěžující přístup.

Stefanowitsch, který zastává profesorský post na Freie Universität Berlin, chtěl podle vlastních slov vytvořit praktickou metodologickou příručku, jakou by on sám jako student býval ocenil, ale na trhu žádná nebyla. S tím nelze než souhlasit: podíváme-li se na repertoár korpusové literatury, najdeme buď obecné úvodní učebnice, které věnují převážnou část teorii a historii korpusů a informace v nich se často překrývají (např. Biber — Conrad — Reppen, 1998; Kennedy, 1998, nebo již zmíněný Hardie — McEnery, 2011), nebo naopak specializované publikace věnující se více či méně přístupně hlavně statistickým otázkám (např. Březina, 2018), často v kombinaci s konkrétními nástroji (např. Gries, 2009; Levshina, 2015; Winter, 2019, všechny věnované statistice s R). Příručka korpusové metodologie, která by postupovala od základů a pečlivě se na příkladech věnovala metodologii a praktické práci s korpusy a s jejich daty, včetně možných úskalí, tu chyběla.

Tato učebnice, jak o ní sám Stefanowitsch důsledně mluví, je rozdělena na dvě pomyslné části. V prvních kapitolách se autor věnuje základním metodologickým konceptům, jako je formulování výzkumné otázky, operacionalizace, odvozování kvantitativních predikcí, extrakce dat a jejich příprava/úprava, statistické vyhodnocení výsledků a jejich interpretace. Ve druhé části, od kapitoly 7 dál, pak rozebírá případové studie z vybraných okruhů lingvistiky (kolokace, morfologie, text, metafora a další), jež byly publikovány v posledních třiceti letech a lze na nich dobře ukázat popisovaný metodologický postup. Knihu doplňuje množství online materiálů, které jsou k dispozici na autorových stránkách nebo v repozitáři na platformě GitHub.

Kdo by se domníval, že už není třeba obhajovat důvody používání korpusových dat, tak by se mýlil — i Stefanowitsch jim věnuje úvodní kapitolu *Need for corpus data* a pregnantně v ní shrnuje hlavní výtky a argumenty proti, protože podle něj mohou

dobře poukázat na ty aspekty korpusového výzkumu, které je třeba si pohlídat. Srozumitelně se vypořádává s námitkami typu, že korpusová data jsou jen úzus, že jsou neúplná, že zahrnují jen formu, a nikoli význam atd. Jasně vyvrací teze považující za lepší zdroj dat intuici, ale zároveň ji ve výzkumu nezatracuje, pouze jí vymezuje patřičný prostor a důrazně doporučuje nezaměňovat ji s daty. Vše doplňuje srozumitelnými příklady, takže i lingvistickou historií nepolíbený čtenář získá solidní přehled o důležitých protiargumentech korpusového výzkumu.

Podobně není očividně možné se obejít bez definice korpusové lingvistiky, hlavně s ohledem na to, že názory na její status se i mezi lingvisty stále různí (velmi zjednodušeně bychom asi mohli definovat dva hlavní tábory: jeden se domnívá, že korpusová lingvistika je samostatná disciplína s vlastní agendou, druhý ji považuje spíše za funkční metodologii použitelnou v mnoha lingvistických disciplínách). Stefanowitsch provádí čtenáře poctivě celým procesem hledání vlastní definice, přičemž vychází z toho, že korpusová lingvistika je vědecká metoda, která zkoumá jazyk na základě korpusu. Na tomto místě pak zdůrazňuje základní vlastnosti korpusu (ovšem opět jen stručně a do té míry, aby mohl pokračovat ve výkladu), jako je autenticita, reprezentativnost vůči zvolené výzkumné otázce či velikost dat. K poslední jmenované uvádí velmi trefně, že jedinou správnou odpovědí na věčnou otázku, jak velký by měl korpus být, je „To není možné říct". K obřím webovým korpusům, jejichž jedinou devizou je právě jejich velikost, se tak autor staví velmi rezervovaně. Čtenář v této kapitole ocení i doplňující a velmi jasné vymezení pojmů žánr, registr, styl či téma (s. 28–29), které mohou být pro mnohé matoucí.

Stefanowitsch se ke své definici korpusové lingvistiky dostává asi na pátý pokus a jedna z finálních verzí, která mně osobně přijde nejpřístupnější, zní: *Corpus linguistics is the investigation of linguistic research questions based on the complete and systematic analysis of the distribution of linguistic phenomena in a linguistic corpus* (tj. korpusová lingvistika spočívá ve zkoumání lingvistických výzkumných otázek na základě kompletní a systematické analýzy distribuce lingvistických jevů v jazykovém korpusu). Vymezením korpusové lingvistiky jako vědecké metody se Stefanowitsch dostává k nezbytnému prvnímu kroku: postulování vědecké hypotézy pomocí dedukce. Vysvětluje postup při formulování hypotézy a zdůrazňuje, že k jejímu zamítnutí nestačí najít pár protipříkladů, ale badatel musí pracovat s pravděpodobností (jak dokládá v následujících kapitolách).

Pojmem, který si zaslouží samostatný odstavec, protože v českých vědeckých publikacích mu mnohdy nebývá věnována zasloužená pozornost, je *operacionalizace*. Tím se myslí postup, při němž se snažíme definovat si zkoumané konstrukty tak, aby bylo možné je najít a ověřit v reálném světě (datech), což je pro empirickou práci naprosto klíčové. Stefanowitsch opět uvádí dostatek příkladů, za všechny např. různě definované slovní druhy nebo délku slova, která může v závislosti na operacionalizaci odpovídat počtu písmen, počtu fonémů nebo třeba počtu slabik. Bez operacionalizace nelze postoupit k dalším bodům solidního korpusového výzkumu, kterými jsou extrakce dat, kvantifikace výzkumné otázky a statistické vyhodnocení dat, jimž autor věnuje následující kapitoly.

Hned v úvodu knihy Stefanowitsch vysvětluje, že z praktických důvodů v učebnici nepoužívá jeden vybraný software, ať už komerční či nekomerční. V podobném

duchu se nese i kapitola o extrakci dat, kde autor zmiňuje různé možnosti a způsoby, včetně programovacích jazyků jako Perl, Python nebo R. Krátce popisuje jen regulární výrazy coby univerzální prostředek, použitelný takřka všude, a uvádí i často používaný jazyk CQL (Corpus Query Language), který je dobře znám i uživatelům Českého národního korpusu. Pozornost věnuje i užitečným pojmům *precision* (přesnost) a *recall* (pokrytí), které srozumitelně vysvětluje na příkladech hned na několika stranách (s. 111–116).

V kapitole o anotaci považuji za klíčové, že se autor více než rozebírání jednotlivých typů anotačních schémat, jak bývá obvyklé, věnuje jejich spolehlivosti a vůbec samotné podstatě anotace. Doslova o ní mluví jako o interpretaci dat, což je zcela zásadní, leč v praxi často opomíjený fakt. Názorně ukazuje, jak různě mohou dva hodnotitelé posoudit tentýž lingvistický jev (s. 132–133) a jak snadno se dá spočítat mezihodnotitelská shoda či reliabilita (*inter-rater reliability*).

Následující dvě kapitoly se věnují tématům, která se mezi mnohými lingvisty netěší přílišné oblibě: kvantifikaci dat a jejich statistickému vyhodnocení. Popravdě není divu, pro humanitně orientované badatele mohou běžné statistické příručky a styl jejich psaní představovat nepřekonatelnou bariéru na cestě ke kvantitativnímu výzkumu (výjimkou jsou uživatelsky přívětivější publikace typu Volín, 2007). U Stefanowitsche ale není třeba se bát, k autorovým přednostem patří srozumitelnost, systematičnost a schopnost nezahltit čtenáře zbytnými detaily a přemírou matematických vzorečků. Základní pojmy jako nominální nebo ordinální data tak díky tomu přestávají být pouhou vzdálenou teorií, ale mění se v konkrétní lingvistické příklady, včetně přehledu vhodných způsobů jejich kvantifikace (např. relativní frekvence, naměřené v. očekávané hodnoty, medián, frekvenční seznam či průměr).

Statistická kapitola je rovněž textem, který bych doporučila k přečtení každému, koho např. termín statistická signifikance děsí (anebo jej naopak zná a představuje vrchol jeho statistického snažení). Stefanowitsch pro každou ze tří kategorií dat, které představil v předchozí kapitole, uvádí jeden vhodný statistický test, který je dle jeho slov natolik jednoduchý, že se dá spočítat s tužkou a papírem nebo kalkulačkou v ruce, a zároveň natolik robustní, aby nenapáchal příliš škody ve chvíli, kdy ho badatel použije špatně (což se, přiznejme si, stává). Jedná se o chí-kvadrát pro nominální data, Man-Whitney U test pro ordinální data a Welchův t-test pro kardinální data. Je jasné, že na mnoho výzkumných otázek se tento výběr hodit nebude, ale jako seznámení s možnostmi statistického vyhodnocení poslouží dobře. Pro ty zkušenější pak autor uvádí i příklady složitějšího výzkumu, v němž figurují proměnné s více než dvěma hodnotami nebo více než dvě proměnné.

Ve zbývajících kapitolách už přicházejí na řadu případové studie z jednotlivých oblastí lingvistického výzkumu, počínaje v korpusové lingvistice oblíbenými kolokacemi. Stefanowitsch neztrácí čas přílišnou teorií, ale důležitým charakteristikám se věnuje pečlivě (včetně přehledu a porovnání nejčastějších asociačních měr — mimochodem, ve svých analýzách v knize pak volí log-likelihood). V rámci tématu se vždy snaží vybrat klíčové a zajímavé studie a ukázat na nich, jak zapadají do metodologického rámce popsaného v první části knihy nebo jak by šlo k problému přistoupit jinak. Příkladem budiž známá studie M. Stubbse z roku 1995 o sémantické prozódii slovesa *cause* (provedená ručně na datech z korpusu LOB), jež Stefanowitschovi slouží

jako východisko k vlastní, jasně popsané analýze na datech z BNC, včetně srovnání tří různých kauzativ a statistického vyhodnocení kolokací pomocí vybrané asociační míry.

Stefanowitsch nevěnuje všem zvoleným oblastem stejnou pozornost, je pochopitelné, že s ním některá témata rezonují víc než jiná. Jen pro představu, z oblasti kolokací a morfologie uvádí v obou případech 6 případových studií, kdežto u metafory a textu jich nabízí 9, respektive 10, a rekordmanem je gramatika, které věnuje hned 13 případových studií s různými tématy.

Z předchozího popisu je jistě zřejmé, že tuto metodologickou příručku považuji za ojedinělý a velmi zdařilý příspěvek do pomyslné korpusové knihovny. Troufám si tvrdit, že se autorovi povedlo to, co si předsevzal — vytvořit učebnici, která na trhu chyběla. Je nabitá informacemi, a přesto si čtenář nepřijde ztracen nebo zahlcen. I proto ji s radostí doporučuji nejen všem svým studentům, nýbrž i kolegům, kteří stejně jako já někdy potřebují spolehlivý maják (a občas i záchranné lano) v bouřlivých vodách korpusových dat.

LITERATURA

Biber, D. — Reppen, R. (eds.) (2015): *The Cambridge Handbook of English Corpus Linguistics*. John Benjamins.

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

Březina, V. (2018): *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge University Press.

Gries, S. (2009): *Statistics for Linguistics with R: A Practical Introduction*. De Gruyter Mouton.

Kennedy, G. (1998): *An Introduction to Corpus Linguistics.* Routledge.

Levshina, N. (2015): *How to do Linguistics with R. Data exploration and statistical analysis*. John Benjamins.

McEnery, T. — Hardie, A. (2011): *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Stubbs, M. (1995): Collocations and cultural connotations of common words. *Linguistics and Education*, 7, 4. 379–390.

Volín, J. (2007). *Statistické metody ve fonetickém výzkumu*. Praha: Epocha.

Winter, B. (2019). *Statistics for Linguists: An Introduction using R*. New York and London: Routledge.

Lucie Lukešová | Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy | Panská 7, 110 00 Praha 1
ORCID ID: 0000-0003-1855-7141
lucie.lukesova@ff.cuni.cz