ČASOPIS PRO MODERNÍ FILOLOGII 100, 2018, Č. 2, S. 271–276

OPEN ACCESS

UWE QUASTHOFF, SABINE FIEDLER, ERLA HALLSTEINSDÓTTIR (EDS.): *FREQUENCY DICTIONARY CZECH. FREKVENČNÍ SLOVNÍK ČEŠTINY*

Leipzig: Leipziger Universitätsverlag GmbH, 2017, 109 stran
ISBN 978-3-96023-157-8

V ediční řadě *Frequency Dictionaries Series* nakladatelství Lipské univerzity vyšel loni již 11. svazek těchto slovníků, *Frequency Dictionary Czech / Frekvenční slovník češtiny*. Je dílem stejné autorsko-editorské skupiny (NLP Group) a má stejný formát jako předchozích 10 svazků, které byly publikovány v tomto pořadí: němčina (GER sv. 1), angličtina (ENG sv. 2), islandština (ISLK sv. 3). francouzština (FRA sv. 4), maďarština (HUN sv. 5), esperanto (EPO sv. 6), indonéština (IND sv. 7), ukrajinština (UKR sv. 8), ruština (RUS sv. 9) a vietnamština (VIE sv. 10).

Cílem série je postupně vytvořit frekvenční slovníky pro velký počet různých jazyků na základě srovnatelných frekvenčních dat. Pro mnoho z těchto jazyků je to první a často jediná příležitost, jak získat frekvenční popis jejich lexika. Každý svazek se skládá ze dvou částí, tištěné brožurky a vloženého CD-ROM. Brožurka obsahuje stručný popis frekvenčních dat, strukturu hesel a koncepci slovníku slovních tvarů. Dále je velmi zběžně popsán použitý korpus vzniklý z internetových materiálů, doba sběru a kroky při zpracování dat. Pro češtinu obsahuje korpus novinové texty, texty z české Wikipedie a největší počet tvoří náhodně vybrané internetové texty (podobné složení má nepochybně i většina ostatních jazyků). Webové rozhraní ke korpusům lze nalézt na adrese http://corpora.informatik.uni-leipzig.de. Toto korpusové rozhraní je možná zajímavější než brožurka. Ačkoli jsou možnosti vyhledávání omezené, korpusy jsou dost velké a zobrazují se kolokace slovních tvarů (i graficky!). Kromě toho lze vyhledávat nejfrekventovanější prefixy a sufixy — ovšem automaticky generované.

Český korpus podle autorů obsahuje 198 milionů vět, všechny údaje jsou ovšem velmi přibližné (ca 18 milionů různých slovních tvarů; slovní tvar byl do slovníku zařazen, pokud se ve zdrojích vyskytl alespoň 33×). Následuje popis statistických dat získaných z vytvořených seznamů slov (včetně statistických údajů o souhláskách, samohláskách a slabikách, počtu písmenných bigramů, trigramů atd.). Tři čtvrtiny brožurky pak zabírají (a) seznam 1 000 nejčastějších slovních tvarů od nejfrekventovanějšího (tj. tvarů „ordered by rank" 1–1000 s uvedením frekvence v ‰ a jejich frekvenční třídy), (b) seznam 10 000 nejfrekventovanějších slovních tvarů seřazených abecedně (s uvedením frekvenční třídy; autoři pracují s 22 frekvenčními třídami vymezenými na základě frekvence daného slova ve vztahu k frekvenci nejfrekventovanějšího slova v korpusu — spojky *a*; první třída 0 obsahuje 2 slova, poslední třída 21 zahrnuje 259 289 slov).

CD-ROM tvořící dodatek k brožurce obsahuje seznam slov v rozsahu 1 000 000 slovních tvarů. Pro češtinu najdeme na CD-ROMu tyto soubory: seznam 1 000 000 tvarů seřazených podle frekvence ve formátu pdf (ve třech dílech) a pak tentýž soubor 1 000 000 tvarů v podobě tří seznamů ve formátu prostý text seřazených (a) podle frekvence, (b) podle abecedy a (c) podle abecedy pozpátku (a tergo). Editory seznamů na CD-ROM byli František Čermák (který přispěl i popisem existujících frekvenčních seznamů a slovníků v češtině v tištěné brožuře), Dirk Goldhahn, Uwe Quasthoff a Maciej Sumalvico.

Je třeba říci, že lipské *Frequency Dictionaries*, které jsou zpracovávány téměř výhradně automaticky, pracují pouze se slovními tvary a neprovádí se lemmatizace. Znamená to, že jedno a totéž slovo-lemma se v tomtéž seznamu objevuje v různých tvarech a zvlášť s malým a velkým písmenem na začátku — pro představu, česká spojka *i* má ve frekvenčním seznamu pořadí (rank) 14, ale psána s velkým písmenem *I* se v seznamu objevuje znovu, ale na 189 místě v pořadí. Jinak řečeno, z uváděných frekvenčních seznamů slovních tvarů se frekvence a počet lemmat dá jen těžko odhadnout.

Z hlediska využití frekvenčních seznamů to samozřejmě může přinášet jistá omezení a nevýhody (např. chceme-li pracovat se slovní zásobou v určitém rozsahu ve výuce nebo sestavovat slovníkový heslář). V případě malých nebo i větších jazyků, kde chybí know-how či peníze na vytvoření korpusu a kde nejsou k dispozici jakékoli frekvenční údaje (tím spíše údaje založené na tak rozsáhlém a recentním materiálu), představují lipské frekvenční slovníky i tak velký přínos. U jazyků, jako je angličtina nebo čeština, pro které jsou již sofistikované frekvenční korpusové údaje k dispozici, budou hrát lipské frekvenční slovníky spíše jen doplňující roli. Přesto podaří-li se sestavit frekvenční slovníky standardizovanou metodou na srovnatelném materiálu pro opravdu velký okruh jazyků, může to být do budoucna podnětem pro zajímavý srovnávací výzkum.

Markéta Malá | Ústav anglického jazyka a didaktiky, Filozofická fakulta Univerzity Karlovy | nám. Jana Palacha 2, 116 38 Praha 1
ORCID ID: 0000-0003-3611-8433
marketa.mala@ff.cuni.cz