RECENZIE

# FREKVENČNÁ ANALÝZA HOVORENEJ ČEŠTINY

***Katarína Gajdošová***
*Slovenský národný korpus Jazykovedného ústavu Ľudovíta Štúra SAV Panská 26, 813 64 Bratislava e-mail: katarinag@korpus.sk*

Jitka ŠONKOVÁ: Morfologie mluvené češtiny: Frekvenční analýza. Praha: Nakladatelství Lidové noviny – Ústav Českého národního korpusu 2008. 356 s. ISBN 978-80-7106-956-0

Šiestym zväzkom edície Studie z korpusové lingvistiky z dielne Ústavu Českého národného korpusu FF UK v Prahe roztvorila vejár počítačových a matematických možností analýzy hovorenej češtiny Jitka Šonková. Už v roku 1995 vo svojej kandidátskej práci Lingvistické zpracování jazyka na počítači: morfologie mluvené češtiny I, II analyzovala a vyhodnocovala materiál približne 40 tisíc textových jednotiek, aby podala exaktný obraz o hovorenej češtine, ktorá z tohto pohľadu ešte nebola skúmaná. Publikácia Morfologie mluvené češtiny: Frekvenční analýza je rozšírením autorkinho výskumu z deväťdesiatych rokov na materiáli Pražského mluveného korpusu (ďalej PMK), ktorý obsahoval v čase prípravy monografie 780 tisíc textových jednotiek.

Monografia je rozdelená na 11 častí. Nezanedbateľnú časť predstavujú prílohy, ktoré tvoria takmer polovicu publikácie (175 strán z celkových 356 strán). V prílohách sa nachádza Kódovník[1], ktorý bol použitý na kódovanie jednotlivých lexém v PMK pri príprave Frekvenčného slovníka mluvené češtiny (2007), a sedem slovníkov lem podľa rôznych parametrov, obsahujúcich frekvenčné ukazovatele ako napríklad absolútny počet výskytov, percentuálny pomer používania a celkový počet výskytov.

V úvode J. Šonková prehľadne predstavila diskusiu českých lingvistov o spisovnej, obecnej a hovorovej češtine spolu s odkazmi na relevantné práce týkajúce sa rozvrstvenia českého jazyka. Autorka prechádza chronologicky problematikou, ktorej východiská a závery (relatívne pre to-ktoré obdobie) sa formulovali už od 30. rokov minulého storočia medzi členmi Pražského lingvistického krúžku (napr. B. Havránek, P. Sgall). Sumarizuje definície a náplň termínu obecná čeština a hovo-

[1] Kódovník je manuál na označkovanie morfologických kategórií jednotlivých slovných tvarov. Zvolené kategórie vychádzajú z morfologických a syntaktických (najmä valenčných) kategórií, ako sú rozpracované v Mluvnici češtiny 2 (1986) a Mluvnici češtiny 3 (1987). Jednotlivé morfologické a syntaktické kategórie sú zachytené vo forme pozičného niekoľkomiestneho číselného kódu.

rová čeština a ich vzťah (B. Havránek, V. Mathesius, J. Bělič, P. Sgall a i.), resp. vymedzenie miesta pôsobnosti týchto variet a ich existenciálnu ne/prítomnosť v českom jazyku. Odkazuje na články a publikácie, v ktorých prebiehali spomínané diskusie a výmeny názorov na túto (v danom čase) kontroverznú tému. J. Šonková nezostáva len pri názoroch českých lingvistov, ale odvoláva sa aj na inojazyčných zahraničných odborníkov v oblasti hovoreného jazyka, najmä angličtiny (J. Blankenship, A. Zettersen, G. Brown – G. Yule). Po predstavení genézy situácie v teórii českého jazyka prechádza autorka k opisu anglických písaných korpusov vo svete. Spomína v 60. rokoch utvorený americký písaný korpus Brown Corpus of Written American English, ale aj jeho konkurenčný korpus britskej angličtiny Lancaster-Oslo/Bergen Corpus of British English. Prehľadne spracovaný vývin korpusov hovorenej angličtiny (americkej a britskej) s podčiarknutím tých najvýznamnejších z nich je aj s odstupom času dôležitým zdrojom informácií (internetové odkazy na citované korpusy sú uvedené v zozname použitej literatúry) najmä pre čitateľov, ktorí korpusy hovorenej angličtiny nepoznajú, resp. sa s problematikou hovorených korpusov iba zoznamujú. V úvode tiež nachádzame informácie o vzniku Českého národného korpusu, ale aj niekoľko slov o budovaní PMK. Čitateľa autorka oboznamuje so svojím predchádzajúcim projektom, v ktorom v 90. rokoch 20. storočia na materiáli subkorpusu pripravovaného PMK (približne 40 tisíc textových jednotiek) skúmala morfológiu hovorenej češtiny na báze frekvenčných ukazovateľov. Práca preverila koncepciu pripravovaného väčšieho výskumu hovorenej češtiny, metódu spracovania, ako aj to, či je zhromažďovaný materiál vhodný pre databázu PMK, a výrazne pomohla pri jeho ďalšom budovaní a neskôr aj pri budovaní Brnenského mluveného korpusu.

S rozšírením PMK a skvalitnením počítačových možností sa J. Šonková rozhodla zopakovať výskum morfológie hovorenej češtiny na ukončenom PMK (cca 780 tisíc textových jednotiek). Využila pritom aj pôvodný Kódovník, aby tak verifikovala a rozšírila závery svojho čiastkového výskumu spred 10 rokov. V druhej kapitole Metoda zpracování jazykového materiálu PMK (s. 20 – 30) má čitateľ k dispozícii podrobnú informáciu o PMK. Autorka sa venuje faktorom, ktoré je nevyhnutné zvážiť pri tvorbe hovoreného korpusu, ako sú napríklad sociologické parametre výberu respondentov, spôsob nahrávania, zachytenie materiálu do textovej podoby a kódovanie slov. Každý respondent má priradený kód, ktorý obsahuje 4 premenné: rod, vek, vzdelanie, komunikačná situácia. Pri rode sa uvádzajú štandardné kategórie muž alebo žena a vekovo sú respondenti rozdelení do dvoch kategórií. Mladší respondenti v rozpätí od 18 do 35 rokov a starší od 36 do 65 rokov. Treťou premennou je vzdelanie a hovoriaci sú rozdelení na hovoriacich so stredoškolským vzdelaním (do tejto skupiny patria aj učňovské odbory s maturitou) a hovoriacich s vysokoškolským vzdelaním. Poslednou zložkou je komunikačná situácia, ktorá sa

označuje ako spontánna alebo riadená. Všetci respondenti boli nahrávaní aj formou riadeného rozhovoru, aj v spontánnej komunikačnej situácii. Formálny rozhovor sa skladal z 12 otázok a odpovedí na ne. Išlo o vyjadrenie názorových postojov respondentov k istým spoločenským, školským či pracovným premisám (napr. *Dnes chodí do práce hodně žen. Je podle vás postavení žen u vás na pracovišti rovnoprávný s muži?*). Prepis nahrávok vychádzal zo zásady, že „všechny tvary slov, které nemají varianty, byly přepisovány spisovně, i když byly vysloveny jinak. Tvary a podoby, které u různých mluvčích alternativy mít mohou, byly zaznamenány přesne tak, jak byly vysloveny“ (s. 23). Každému slovnému tvaru bola pri prepise pridelená lema, k textovému tvaru bola priradená tzv. štýlová a spisovná lema (napr. *takovym* – upravený tvar z prepisu, *takovej* – štýlová lema, *takový* – spisovná lema). Štýl tvoria v PMK štyri triedy: 1) nespisovné tvary, v jazyku používané ako hovorené a neformálne a kódované ako základné; 2) spisovné tvary, používané ako písané a hovorené, formálne aj neformálne a kódované ako neutrálne; 3) knižné tvary, v jazyku používané ako písané a formalizované; 4) vulgárne tvary, v jazyku používané ako hovorené a neformalizované.[2]

Rozboru morfologických kategórií slovných druhov v hovorenej češtine predchádza štatistická analýza PMK ako celku zo sociolingvistického hľadiska. Analýza frekvencie slovných druhov bola urobená pre PMK podľa štýlového indexu spolu s charakteristikou najfrekventovanejších spisovných lem, štýlových lem a slovných tvarov. Autorka zdôrazňuje, že výsledky, ktoré vyplynuli z analýzy PMK podľa štýlového indexu so zreteľom na sociolingvistické skupiny, boli jedným z hlavných motivačných faktorov výskumu. Prostriedky knižného a vulgárneho štýlu sú zanedbateľné a ďalej s nimi autorka nepracuje (nepresiahli 0,1 %). Zaujímavá je ale odchýlka od relatívnej frekvencie jednotlivých sociolingvistických kategórií v rámci hovoreného neformálneho a formálneho štýlu. Z čísel v tabuľkách sa dozvedáme, že „styl promluvy je nejvíce ovlyvňován typem rozhovoru, následuje vzdělání a věk. Za zmínku stojí skutečnost, že vyšší vzdělání mluvčích je spojeno s častějším výskytem slovních tvarů neformálního stylu. Mluvčí s vyšším vzděláním užívali nespisovné tvary častěji (93,4 %) než mluvčí s nižším vzdeláním (85,4 %). Mluvčí v řízených rozhovorech, ve ktorých odpovídali na výše uvedené otázky, užívali o 11,6 % více

[2] V Slovenskom hovorenom korpuse (ďalej SHK) je každému hovoriacemu takisto pridelený kód zložený z písmen, ale sociolingvistické informácie týkajúce sa hovoriacich sú širšie ako v PMK. Okrem štandardného delenia respondentov na mužov a ženy zachytáva SHK aj informácie o ich vzdelaní v štyroch kategóriách – vysokoškolské, stredoškolské, odborné (učňovské odbory) a nižšie. Metadáta tiež obsahujú záznam o mieste narodenia respondenta, mieste jeho najdlhšieho pobytu a mieste jeho súčasného pobytu. Vek je škálovaný od veku 20 rokov po dekádach, pričom respondent môže uviesť aj presný vek. Nechýba ani informácia o materinskom jazyku respondenta, ďalších jazykoch, ktoré aktívne ovláda, informácia o používaní nárečia (v prípade súhlasnej odpovede aj jeho presnú špecifikáciu) a najdlhšie vykonávaného povolania. Podrobnejšie informácie o SHK sú zverejnené na webovej stránke projektu http://korpus.sk/shk.html.

formálních (spisovných) výrazů než ve spontánních rozhovorech“ (s. 40). Ako všetky nasledujúce kapitoly aj táto obsahuje mnoho graficky zobrazených štatistických údajov vzťahujúcich sa na jednotlivé hodnoty. Autorka uvádza relatívne frekvencie slovných tvarov podľa jednotlivých sociologických parametrov (rod, vek, vzdelanie, typ rozhovoru), v grafoch zachytené absolútne a relatívne frekvencie slovných druhov a slovných tvarov podľa štýlového indexu a v tabuľke napríklad zoznam prvých tridsiatich najfrekventovanejších slov, ktoré porovnáva podľa spisovnej lemy, štýlovej lemy a slovného tvaru. Konštatuje, že prvé dva parametre porovnávania sa do desiatej pozície líšia len usporiadaním lexém. Frekvenčné hodnoty často sprevádzajú vzorce a výpočty, ktoré sú príznačné pre matematickú lingvistiku. Pri znázornení relatívnej frekvencie prvých štrnástich spisovných lem môžeme vidieť, akou rýchlosťou klesá relatívna frekvencia od najfrekventovanejšej lemy *ten* až po 14. najfrekventovanejšiu lemu *tedy*. Matematicky je nutné vzorcom zachytiť korelačný koeficient. S jeho použitím pri prepočtoch autorka znázorňuje krivkou v grafe porovnanie relatívnej frekvencie prvej stovky najfrekventovanejších spisovných lem. Ide o vynikajúcu zhodu ($R^2$=0.99) medzi priebehom funkcie a algebrickou krivkou grafu[3].

V ďalších kapitolách J. Šonková podrobne uvádza frekvencie prvých piatich slovných druhov a číselné hodnoty týkajúce sa ich podkategórií tak, ako boli označkované na základe Kódovníka. Napríklad na prvom mieste sa nachádza slovný druh (s číslom 1 substantíva), na druhom mieste[4] druh (bežné, adjektívne, zámenné, číslovkové, slovesné, slovesné zvratné, skratkové slovo, nesklonné substantívum). Tretie miesto patrí triede (osoba, živočích, konkrétna vec, abstraktná vec, iné/nejasné), na štvrtom mieste je valencia (bez valencie, s bezpredložkovým pádom, s predložkou, so spojovacím výrazom, s infinitívom, s adverbiom, s dvoma bezpredložkovými pádmi, s bezpredložkovým a predložkovým pádom, s bezpredložkovým pádom a spojkou, iné a viacpočetné). Piatou substantívnou kategóriou v poradí je rod (mužské životné, mužské neživotné, ženské, stredné, nie je možné určiť). Šiesta pozícia v číselnom kóde patrí číslu (singulár, plurál, plurál tantum, kolektívum, nie je možné určiť), siedma pozícia pádu (nominatív, genitív, datív, akuzatív, vokatív, lokál, inštrumentál, nie je možné určiť). Na ôsmej pozícii je zachytená funkcia (subjekt, predikát, nezhodný atribút, neväzobné adverbiále (aj s predložkou), vokatívna veta, nominatívna veta, iná veta, iné, samostatné, nie je možné určiť). Svoje špecifické podkategórie majú všetky skúmané slovné druhy, pričom každej podkategórii je priradené na istej pozícii v číselnom kóde vlastné číslo. Text jednotlivých frekvenčných analýz slovných druhov obsahuje mnoho tabuliek, grafov rôzneho druhu a konkor-

[3] Hodnota $R^2$=0 ukazuje, že medzi algebrickou krivkou a skutočným priebehom funkcie nie je žiadna závislosť, $R^2$=1 naopak hovorí o úplnej zhode.

[4] Uvádzané podkategórie sa vzťahujú na slovný druh substantíva.

dančných výstupov z korpusu. Autorka zdôrazňuje: „Jádrem každé sekce zabývající se ohebnými druhy slov je frekvenční analýza rozdílů v běžně mluveném a spisovném jazyce doplněná konkordancemi nejčastejších tvarů“ (s. 50). V kapitole Diskusia výsledkov (s. 158 - 170) autorka opätovne zhŕňa najdôležitejšie štatistické ukazovatele jednotlivých slovných druhov, ako ich obsahuje PMK, a porovnáva frekvenčné výsledky s frekvenčnými ukazovateľmi tých istých slovných druhov z písaného korpusu češtiny SYN2005. Z dvadsiatich najfrekventovanejších lexém v PMK a SYN2005 sa jedenásť zhoduje, z prvých šiestich sú tri zhodné (*ten, být, a*), líšia sa však poradím. Z hľadiska porovnania slovných druhov sú v PMK najfrekventovanejšie slovesá (18 %), zámená (17 %), častice a substantíva (po 13 %), v SYN2005 sú to substantíva (31 %), slovesá (17 %) a adjektíva (12 %). J. Šonková vysvetľuje: „...tento rozdíl je zcela logický vzhledem k různé funkci analyzovaných projevů. Pro mluvený jazyk je charakteristická dějovost (slovesa), opisování a substituce, zvláště jmen, mimo jiné i orientace v časoprostoru (zájmena) a vyjadřování postoje, názoru a emocí (částice)“ (s. 159).

V závere monografie autorka konštatuje, že hovorená čeština je jazykom, ktorý v rozličnom pomere obsahuje prvky spoločné pre spisovný aj nespisovný jazyk. Z porovnania pôvodného výskumu hovorenej češtiny na báze 40 tisíc slov a opakovaného výskumu na mnohonásobne väčšom PMK vyplynulo, že rozdiely sú percentuálne zanedbateľné a aj výskum hovorenej češtiny na rozsiahlom materiáli potvrdil pôvodné závery, že hovorená čeština „je konzistentní, ustálený a pravidla zachovávající jazykový systém, jehož pravidla se však poněkud liší od kodifikované spisovné češtiny“ (s. 171). Pri rozhodovaní sa o použití spisovného alebo nespisovného tvaru v prípade variantov si českí hovoriaci v 75 % prípadov vyberajú nespisovný tvar. Pri hlbšej analýze frekvencie nespisovných variantov sa ukázalo, že boli zväčša použité v kontexte iných nespisovných morfologických tvarov typických pre nepripravené neformálne prejavy. Výsledky frekvenčnej analýzy PMK, a teda všeobecne českých hovorených prejavov, môžu byť perspektívne využité napríklad pre hovoriacich, ktorí ako cudzinci študujú češtinu, alebo ako odrazový mostík pri tvorbe učebníc pre cudzincov, ale aj ako dôležitý zdroj pre lingvistov na ďalší výskum hovorenej češtiny.

Publikácia J. Šonkovej Morfologie mluvené češtiny: Frekvenční analýza je skutočne dôkladným štatistickým rozborom morfológie hovorenej češtiny. Zosumarizované grafy, tabuľky a konkordančné výstupy predstavujú cenné východisko na ďalšie výskumy hovorenej češtiny. Autorka si nekládla za cieľ vysvetliť špecifické javy hovorenej češtiny, a preto jednotlivosti nájdeme v publikácii len veľmi okrajovo. Jej cieľom bolo skôr podať komplexný pohľad na hovorený jazyk ako taký a ukázať, že hovorená čeština je vo svojej podstate samostatný jazyk ustálený svojou formou. Na jeho ďalší opis slúži aj materiál zhromaždený a roztriedený v jej monografii.

Literatúra

Český národní korpus – PMK. Praha: Ústav Českého národního korpusu 2001. Dostupný z WWW: korpus.cz
Český národní korpus – SYN2005. Praha: Ústav Českého národního korpusu 2005. Dostupný z WWW: korpus.cz
Frekvenční slovník mluvené češtiny. Ed. F. Čermák. Praha: Karolinum 2007. 510 s.
Mluvnice češtiny 2. Praha: Academia 1986. 536 s.
Mluvnice češtiny 3. Praha: Academia 1987. 746 s.
Slovenský hovorený korpus. Informácie o projekte dostupné na WWW: http://korpus.sk/shk.html.
ŠONKOVÁ, Jitka: Lingvistické zpracování jazyka na počítači: morfologie mluvené češtiny I, II. Kandidátska disertační práce. Praha: Filozofická fakulta UK 1995.

# NETRADIČNÁ MONOGRAFIA O SÚČASNEJ ČESKEJ TOPONYMII

***Lucia Luptáková***
*Katedra slovenského jazyka a literatúry s oddelením prekladateľstva a tlmočníctva, Fakulta humanitných vied Univerzity Mateja Bela, Tajovského 4, 974 01 Banská Bystrica*
*e-mail: lucia.luptakova@umb.sk*

(David, Jaroslav: Smrdov, Brežněves a Rychlonožkova ulice. Kapitoly z moderní české toponymie. Praha: Academia 2011. 333 s. ISBN 978-80-200-1913-4)

V roku 2011 rozšírila onomastický knižný trh monografia o modernej českej toponymii Smrdov, Brežněves a Rychlonožkova ulice. Jej autor, český onomastik a diachrónny lingvista, Jaroslav David, prestavil knihu slovami: „Smrdov, Brežněves a Rychlonožkova ulice s podtitulem Kapitoly z moderní české toponymie představuje nový a netradiční pohled na vývoj a preměny soustavy místnich jmen ve 20. století a svým tématem je tedy prvním pokusem o zpracování této problematiky nejen v české, ale i středoevropské toponomastice.“ (s. 17). Závan novosti a netradičného spracovania predznamenáva nielen názov, ale i tematické spracovanie jednotlivých kapitol, ktoré na pozadí ústredného motívu honorifikácie podávajú výklad súčasnej českej toponymie.

Honorifikáciu vysvetľuje J. David na základe pozorovania oslabujúceho sa vzťahu medzi propriom a označovaným objektom. Honorifikačné meno neplní základné onymické funkcie (identifikácia, diferenciácia, lokalizácia), ale vyjadruje úctu alebo vďačnosť nejakej osobe, t. j. honorifikáciu chápe ako nepriamu posesivitu. Motívom pomenovania tu nie je vlastnícky vzťah osoby, podľa ktorej je miesto po-