OPEN ACCESS

VÁCLAV CVRČEK, ZUZANA LAUBEOVÁ, DAVID LUKEŠ, PETRA POUKAROVÁ, ANNA ŘEHOŘKOVÁ, ADRIAN JAN ZASINA: *REGISTRY V ČEŠTINĚ*
Praha: Nakladatelství Lidové noviny, 2020, 227 stran
ISBN 978-80-7422-754-7

Jazyková variabilita je inherentním rysem jazyka, který prostupuje řadu oblastí jazykového zkoumání. Již od sedmdesátých let poukazují někteří badatelé na to, že žádný mluvčí nemluví ve všech situacích stejně a variabilita jazykových rysů (a dále pak i textů) je podmíněna situačně a funkčně (např. Halliday — Hasan, 1976; Hymes, 1974). Mluvčí musí při formulaci textu činit rozhodnutí, která jsou podmíněna funkcí textu, tématem, vztahem mezi mluvčím a adresátem, mírou interakce a situačním zakotvením projevu či textu. Metodu, která umožňuje rozplést tuto souvztažnost a jazykovou komplexitu, definoval a poprvé aplikoval Douglas Biber ve své studii *Variation in Speech and Writing* (Biber, 1988) a posléze ji využil v mnoha dalších výzkumných projektech. Jeho badatelská práce se stala vzorem a inspirací i pro autorský tým této publikace. Autoři si kladou za cíl empiricky a metodicky rozšířit popis variability v češtině a vyvinout vůbec první multidimenzionální model slovanského jazyka. Jejich monografie přehledně shrnuje postup při vývoji daného modelu a definici registrů češtiny a dále popisuje uplatnění modelu v aplikovaném výzkumu. Publikace je rozdělena do sedmi kapitol, je opatřena přílohou se soupisem výrazných rysů v jednotlivých dimenzích, seznamem použitých rysů a seznamy pro operacionalizaci rysů a jmenným rejstříkem.

První kapitola slouží jako úvod do problematiky jazykové variability, resp. do jejího funkčního využití, objasňuje strukturu knihy a záměry jednotlivých kapitol. Za zmínku stojí i poznámka na konci první kapitoly, která poskytuje odkazy na zdroje a nástroje použité při realizaci MDA.

Druhá kapitola představuje teoretická východiska multidimenzionální analýzy. Aplikovaná metoda vychází ze studie D. Bibera (1988): byla vyvinuta pro analýzu registrů v angličtině a od té doby byla jen s menšími úpravami použita a dále badatelsky ověřena v řadě studií zaměřujících se i na specializované registry. Jádrem Biberovy metody je myšlenka, že jazyková variabilita je podmíněna funkčně a situačně. Jedná se v podstatě o explorativní metodu, která nám umožňuje zkoumat vztah mezi jazykovými rysy a textovými charakteristikami. Při koncipování metody se ale nepostupuje od funkčních vysvětlení (např. ve smyslu škál formální — neformální), ale výzkum je postaven na analýze výskytů, resp. souvýskytů jazykových rysů, které jsou determinovány komunikační situací a funkcí daného textu (s. 13–15).

Studie odkazuje na dosavadní poznatky české stylistiky a empirické lingvistiky, přičemž zmiňuje dvě perspektivy pohledu na text. Perspektivu vnitrotextovou, která vychází z jazykových prostředků použitých v textu s důrazem na to, že jejich výběr neprobíhá izolovaně či jednotlivě, ale jako „seskupení takovýchto jazykových prostředků, které realizují základní komunikační záměr původce jazykového projevu …“ (s. 16). Druhá perspektiva zprostředkovává pohled vnětextový, kde jsou určující faktory *stylu* jako komunikační záměr, ráz komunikace, přítomnost a nepřítomnost adresáta, mluvenost a psanost, resp. připravenost a nepřipravenost. Toto rozlišení umožňuje autorům vymezit pozici registru jako útvaru, „který je rozkročen mezi

vnětextovým a vnitrotextovým pohledem“ (s. 20). To znamená, že v něm zohledněn vztah mezi situačním kontextem, zvolenými jazykovými prostředky a jejich funkčním zapojením v textu.

Třetí kapitola umožňuje čtenáři nahlédnout do motivace autorů a seznámit se s jednotlivými kroky při zpracování multidimenzionálního modelu. Autoři jsou vedeni snahou o poznání toho, „jak texty variují, jaké jsou základní tendence (dimenze variability) a jaké rysy se na tom podílejí“, a dále pak tím, „jak se vliv registru např. promítá do frekvenčních charakteristik různých rysů“ (s. 21.). Výstupy z MDA dále umožňují poznat rozpětí variability českých textů, což je stěžejní v oblasti designu korpusů a při koncepci vnitrotextové klasifikace textů v korpusu.

Následuje detailní a precizní popis pracovního postupu celé analýzy, který zahrnuje sestavení korpusu, kompilaci jazykových rysů pro analýzu, operacionalizaci rysů, statistické vyhodnocení a interpretaci výsledků. Samotná koncepce a sestavení korpusu zahrnuje několik kroků a rozhodnutí, které nelze z prostorových důvodů v této recenzi vyčerpávajícím způsobem popsat (s. 24–33). Podstatné je, že pro účely MDA analýzy bylo třeba sestavit korpus, který by zachycoval šíři a pestrost variability, přičemž reprezentativnost ve vztahu k populaci (celému jazyku — zde češtině) nebyla zas tak důležitá. Za účelem dosažení pestrosti bylo do korpusu zařazeno 3428 vzorků, přibližně 10,9 milionů tokenů. Vybrané texty pokrývají tři komunikační módy (psaný/*wri*, mluvený/*spo* a internet /*web*), které jsou dále členěny na divize a (nad)třídy, přičemž každá třída obsahuje přibližně 200 tisíc tokenů. Autoři zdůvodňují užití vzorků potřebou zajistit vyhraněnost textů, a tím i jejich jasnou pozici v rámci prostoru variability. Tato homogenita totiž u delších textů zpravidla chybí. V podstatě lze říci, že čím delší text, tím větší prostor pro různé jazykové prostředky a tím větší různorodost textu, a tudíž menší vyhraněnost.

Druhým krokem v pracovním postupu MDA byl výběr jazykových rysů na různých jazykových úrovních (fonologické, morfologické, slovotvorné, lexikální, syntaktické, pragmatické i textové), které reprezentují variabilitu češtiny. Seznam rysů má základ v mluvnicích a stylistických příručkách češtiny, vybraných úzce zaměřených studiích, a rovněž v úvahách autorů, v nichž hraje nespornou roli i jazyková intuice. Při operacionalizaci (tj. formulaci korpusového dotazu a automatické extrakci frekvencí daných jevů) se autoři snaží zajistit, aby byl do výsledku zahrnut daný rys v co největší šíři (maximálně možný počet relevantních případů) s co nejvyšší přesností (s minimálním počtem chyb při extrakci). Na základě těchto hledisek byly z analýzy vyřazeny rysy s velmi nízkou frekvencí, vysokou chybovostí a minimálním podílem na variabilitě (např. gramatický rod). V podkapitole 3.4.2 (s. 36–77) je uveden spíše technický popis toho, jak byly jednotlivé rysy extrahovány. Tato část i přes svou techničnost má svůj význam, a to hned z několika důvodů. Za prvé slouží jako přesná a detailní dokumentace ve smyslu dobré badatelské praxe. Za druhé umožňuje čtenářům, kteří se s kvantitativní a korpusovou lingvistikou spíše seznamují, nahlédnout do samého nitra práce s jazykovým korpusem. Za třetí, pro čtenáře, kteří s korpusovým bádáním jisté zkušenosti mají, může tento detailní popis sloužit jako inspirativní zdroj při koncepci vlastních výzkumných projektů.

Podkapitola 3.5 *Statistické vyhodnocení* podává zdůvodnění výběru faktorové analýzy mezi metodami na redukci dimenzionality, osvětluje její základní principy

a demonstruje krok za krokem její aplikaci. Faktorová analýza předpokládá, že mezi analyzovanými (jazykovými) rysy existují korelace, a tudíž není nutno je analyzovat jednotlivě, nýbrž jako „balíček" korelujících rysů, jejichž souvýskyt je podmiňován latentní proměnnou, která se projevuje jako faktor. Vstupem analýzy byly hodnoty (zpravidla frekvence) pro 122 jazykových rysů pro 3292 textů, které byly upraveny a normalizovány a bylo přistoupeno k extrakci faktorů. Podkapitola dále zdůvodňuje metodicky klíčová rozhodnutí týkající se počtu extrahovaných faktorů, typu rotace (*promax*) a metody extrakce faktorů (GLS). Výstupem faktorové analýzy jsou dva typy informací, tzv. *loading*, který informuje o zapojenosti rysu v daném faktoru (či dimenzi) a *factor score*, což „je charakteristika textu, která je odvozena od [jazykových] rysů v něm použitých". Bude-li text obsahovat hodně rysů s pozitivním loadingem a málo či žádné s *loadingem* negativním, bude mít vysoké *factor score,* a naopak (s. 82). Tato data jsou zásadní pro interpretaci dimenzí, která je shrnuta v podkapitole 3.5.5. Autoři zde dávají nahlédnout do procesu posuzování metod užívaných k určování počtu dimenzí, jasně vysvětlují podmínky a problémy aplikace, které neumožňují jejich využití u českého modelu. Dále detailně popisují (s. 84–90) a odůvodňují vlastní postup při interpretaci dimenzí. Celý postup od kalkulace přes implementaci je dostupný k nahlédnutí v repozitáři *Github*.

Ve čtvrté kapitole je krátce nastíněn způsob interpretace (korelujících) rysů ve vztahu k jednotlivým faktorům a jsou zde určeny příslušné krajní hodnoty a inertní rysy pro daný faktor (resp. dimenzi). Je prozkoumána variabilita subkorpusů *Koditexu* a příslušné rozdíly mezi nimi. Výsledky těchto srovnání lze také prostudovat ve vizualizačním nástroji vyvinutém pro účely MDA analýzy, který je volně dostupný na stranách Českého národního korpusu (www. korpus.cz/mda). Následuje podrobný popis dimenzí variability, který obsahuje jejich typické rysy, textové typy významné v dané dimenzi a ilustrativní ukázky daných textů. Podkapitola 4.2 *Srovnání s Biberovým modelem* se zajisté lépe čte informovanému čtenáři, který automaticky asociuje Biberův model, nicméně i pro ty méně informované jsou cenné poznatky ohledně specifičnosti českého MDA modelu. Zajímavé je například vyčlenění 2. dimenze spontánní (+) vs. připravený (–) či 4. dimenze polytematický (+) vs. monotematický (–). V další podkapitole (4.3) se autoři vrací k otázce vlivu délky vzorků (chunků) pro stanovení modelu češtiny. Na příkladu s webovými daty ukazují, že spolehlivost zařazení do dimenze je problematičtější v dimenzích, na kterých se podílejí méně časté jazykové rysy. Zatímco dimenze, na kterých se podílejí obecně časté rysy (frekvence substantiv, adjektiv, verb), jsou robustnější, a tudíž ke zkreslení způsobené délkou vzorku méně náchylné (s. 114–115).

I když jsou dimenze pro popis variability díky své škálovosti a hierarchičnosti nesporně přínosné, dle autorů má jazykový popis prostřednictvím dimenzí několik nevýhod: zpracování osmi dimenzí se pohybuje na hranici kognitivních možností a v určitých oblastech jako např. klasifikace textu je jedno (jediné) označení žádoucí (s. 116). V páté kapitole se tedy autoři dostávají k záměru identifikace registrů češtiny, tzn. k vnitrotextově motivovanému popisu skupin (shluku) textů. Čtenář si jistě může klást otázku, k čemu takový popis slouží. Ve prospěch identifikace registrů hovoří, kromě výše zmíněných nevýhod dimenzí, i to, že shluknou-li se texty do skupin, lze vymezit jejich pozici a velikost (obecnost či specifičnost) a určit jejich homogenitu

(s. 117). Pro identifikaci klastrů byla použita klastrovací metoda KMeans, počet klastrů byl určen pomocí balíku *NbClust* (R Core Team 2018) a jako nejvhodnější se ukázala metoda deseti klastrů, tj. 10 registrů. Přehled registrů je popsán v podkapitole 5.3. Pro každý registr je uvedeno jeho jméno, skládající se ze dvou částí. První je inspirována slohovým postupem, formátem či záměrem (s. 119), druhá část se vztahuje k pozici klastru v první dimenzi (dynamický vs. statický) a druhý atribut k dalšímu výraznému rysu klastru v další dimenzi (např. analýza: statický monotematický registr, komentář: dynamický postojový registr). Popis registrů nám podává informace o tom, kde se registr v daných dimenzích nachází, jaké jazykové rysy jsou v něm významně zastoupeny a které typy textů jsou pro tento registr typické. I zde je přiložena textová ukázka.

V šesté kapitole uvádí autoři příklady aplikace multidimenzionálních modelu češtiny ve třech studiích. První studie měla posoudit, zda texty elicitované pro psycholingvistickou analýzu na základě čtyř scénářů dosahují kýžené variability. U dvou scénářů bylo zjištěno, že texty produkované na jejich základě se neodlišují natolik, aby mohly být spolehlivě použity v plánovaném výzkumu. Důležitým poznatkem této studie je, že při výzkumech zohledňujících jak jazykové, tak i psychologické proměnné lze takovouto analýzu vstupů jen doporučit. V druhé studii se autoři zaměřují na to, do jaké míry je variabilita ovlivněná situací (registrem) a do jaké míry idiolektem daného mluvčího. Vstupem jsou výše zmíněné elicitované texty (dopisy). Vyhodnocení probíhalo čtyřmi způsoby: modelováním prostřednictvím tří statistických metod a porovnáváním rozdílů prostřednictvím vzdáleností mezi texty. Použití čtyř metod za stejným účelem umožňuje spolehlivou verifikaci výsledků, tzn. jestliže vyjde pomocí čtyř metod podobný výsledek, tak jsme s největší pravděpodobností postupovali správně. Studie dochází k závěru, že registr se na variabilitě textu podílí významněji než idiolekt (poměr 2:1). Na základě těchto výsledků autoři doporučují, aby studie, které se zaměřují na výzkum idiolektu, pracovaly s texty pouze z jednoho registru, a tím byl omezen jeho vliv na variabilitu. Třetí studie porovnává variabilitu dat z tradičních (reprezentativních) korpusů (zde korpus *Koditex*) a jazykových dat získaných z webu (zde korpus *Araneum Bohemicum Maximum)*. V podstatě jde o odpověď na otázku, zda jsou ve velkém objemu webových dat obsaženy všechny textové typy a zda mohou webová data suplovat data mluvená. Z výsledků vyplývá, že mezi daty z *Koditexu* a z webu je sice velký překryv, ale některá uživatelsky generovaná data jsou zastoupena jen částečně. Ukazuje se také, že běžně crawlovaná data nepokrývají celé spektrum jazykové variability. Podle autorů by při analýze webových dat měli badatelé počítat s tím, že část variability pokrytá např. spontánními mluvenými rozhovory či autorskými texty, např. beletristickými, bude ve webovém korpusu chybět. Upozorňují dále na to, že některé chybějící webové (autorsky generované) textové typy je vhodné doplnit prostřednictvím specializovaných webových korpusů (Facebook, Twitter atd.).

Závěrem lze říci, že tato publikace a s ní spojený výzkumný projekt jsou velkým přínosem k poznání variability v češtině a českých registrů. Daný model nejen splňuje svou deskriptivní funkci, ale také otevírá další perspektivy pro estimaci variability v aplikovaném výzkumu (kap. 6), což výrazně rozšiřuje dosavadní možnosti lingvistického bádání v češtině. V neposlední řadě je třeba vyzdvihnout transparentnost

prezentované studie v oblasti dokumentace a zpřístupnění relevantních dat, skriptů a nástrojů, a ocenit čtivost a srozumitelnost stylu publikace. Tato monografie, a především způsob badatelské práce v ní zachycený může bezesporu sloužit dalším odborníkům jako příklad dobré vědecké praxe.

LITERATURA

Biber, D. (1988): *Variation in Speech and Writing*. Cambridge: Cambridge University Press.

Halliday, M. A. K. — Hasan, R. (1976): *Cohesion in English*. London: Longman.

Hymes, D. (1974): *Foundations in sociolinguistics: An enthnographic approach.* Philadelphia: University of Pennsylvania Press.

Martina Berrocal | Institut für Slawistik und Kaukasusstudien, Friedrich-Schiller-Universität Jena | Ernst-Abbe-Platz 8, 07743 Jena
ORCID ID: 0000-0003-4003-8516
martina.berrocal@uni-jena.de