OPEN ACCESS

VÁLKA S ŽÁNRY

David Lukeš

Kubát, M. (2016). *Kvantitativní analýza žánrů*. Ostrava: Ostravská univerzita, Filozofická fakulta, 141 s.

Miroslav Kubát ve své publikované dizertaci zkoumá paralely mezi konvenčním žánrovým zařazením textu a jeho jazykovými rysy, měřenými pomocí kvantitativně lingvistických metod. Snaží se tedy zodpovědět otázky týkající se vztahu mezi vnětextovou a vnitrotextovou charakteristikou textu: Vyznačuje se báseň vyšší nebo nižší tematickou koncentrací než román, nebo jsou oba žánry srovnatelné? Jak jsou na tom z hlediska lexikální bohatosti? atp. Používá při tom široké spektrum stylometrických metod, v jejichž pečlivém výběru a srozumitelně formulované charakteristice tkví největší přínos celé práce. Nicméně jak vyplyne z následujících řádků, dva klíčové aspekty práce vybízejí přinejmenším k polemice. Jmenovitě jde o volbu materiálu, s nímž Kubát pracuje (jedná se o výbor z díla jediného autora, Karla Čapka), a různé aspekty analýzy dat (statistické vyhodnocení, jeho interpretace, vizualizace výsledků).

Věnujme se však nejprve jednoznačně kladným aspektům práce. Výběr použitých stylometrických indikátorů je podložen pečlivým studiem sekundární literatury, v níž se autor velmi dobře orientuje, jak ostatně dokazuje i fundovaným výkladem zvažovaných metod, rozborem jejich kladů a záporů a názornými příklady. Metody dělí do dvou skupin, přičemž první z nich bychom mohli označit za lingvisticky motivované, a tudíž i interpretovatelné, druhou pak za pragmatické. Autor sám o této druhé skupině hovoří jako o metodách, „které se primárně používají ke klasifikaci textů, zejména pak v oblasti určování autorství" (s. 38), a zařazuje je spíš jako zajímavost pro doplnění. Do první kategorie řadí různé způsoby určování slovního bohatství odvozené od *type-token ratio* (TTR), tematickou koncentraci textu, vzdálenosti sloves, průměrnou délku tokenu a tzv. aktivitu; do druhé pak charakteristiku textu pomocí n-gramů a nejfrekventovanějších slov. Tímto dělením autor vhodně poukazuje na to, že cílem práce není ani tolik hrubou silou co nejlépe automaticky identifikovat žánry, jako spíš empiricky zjistit typické jazykové rysy, které tyto žánry vykazují, a pokusit se je lingvisticky interpretovat.

Nemá smysl zde reprodukovat popisy jednotlivých metod, jak již bylo řečeno, výběr je poučený a výklad přístupný, takže k tomuto účelu nejlépe poslouží publikace samotná. Jako jeden příklad za všechny uveďme pojednání o slovním bohatství v kap. 4.1. Kubát srozumitelně vysvětluje, proč má na surové TTR vliv délka textu, což je pochopitelně pro porovnávání textů různých délek problém, a rozebírá postupy, kterými se různí badatelé snažili tento problém obejít. Přehled je to dobře uspořádaný a čtivý, osobně bych do něj snad jen ještě zařadil metodu zTTR (Cvrček & Chlumská, 2015). Ne snad proto, že by přehled nutně musel být vyčerpávající, ale proto, že tato metoda se od ostatních uvedených zajímavým způsobem liší,[1] takže by obohatila

1 Zatímco ostatní metody pracují jen s cílovým textem samotným, v zTTR se porovnává TTR cílového textu s referenčními hodnotami TTR reprezentativního vzorku textů podobné délky.

už tak pestrou diskusi ohledně možných přístupů k řešení tohoto problému. Na druhou stranu, dle vročení je patrné, že článek o zTTR a Kubátovu publikaci od sebe dělí pouhý rok, takže absenci zTTR lze patrně zčásti přičíst i souběhu obou výzkumů.

Sám Kubát nakonec po přesvědčivé argumentaci volí metodu Moving Average TTR (MATTR), která spočívá ve zprůměrování hodnot TTR vypočtených na základě posouvání okna o stanovené šířce po textu token po tokenu. Nicméně zmiňuje i z ní odvozenou metodu Moving Window TTR Distribution, na jejímž vývoji se sám podílel. V této variantě se TTR naměřená v jednotlivých oknech neprůměrují do jednoho bodového odhadu, ale sleduje se jejich rozdělení, což umožňuje jemnější porovnání, ovšem za cenu složitějšího statistického vyhodnocení. I proto autor pro hlavní analýzu zůstává u MATTR.

Vzhledem k této nuancované a dobře podložené selekci metod mi přijde o to větší škoda, že z hlediska materiálu, na kterém žánrovou variabilitu zkoumá, se Kubát (cíleně a vědomě) omezuje na tvorbu jediného autora (Karla Čapka). Autor nabízí následující vysvětlení:

> Omezením korpusu analyzovaných textů na jediného autora jsme dosáhli eliminace nežádoucího vlivu různých autorských stylů, jenž nutně devalvuje vypovídací hodnotu podobných výzkumů. Každý autor má totiž svůj jedinečný způsob psaní, který prostupuje napříč žánry. V případě zařazení více autorů do korpusu nutně znemožníme relevantní vyhodnocení výsledků, neboť nebudeme s to zjistit, zda získané hodnoty vypovídají spíš o autorovi nebo o žánru. Na druhou stranu je třeba poznamenat, že omezením korpusu na jediného autora nelze závěry zobecnit, protože nevíme, jak by se sledované ukazatele chovaly v případě jiných autorů (s. 11–12).

Kdybychom na tuto argumentaci přistoupili, znamenalo by to, že jakákoli snaha zkoumat žánry je marná. Buď se totiž omezíme na jediného autora a naše výsledky nebudou zobecnitelné, nebo jich budeme zkoumat víc, ale výsledky budou irelevantní, protože se v nich bude mísit vliv žánru a autorského stylu. To je ovšem falešné dilema. Žánr jako takový prostřednictvím jediného autora zkoumat jednoduše nelze, neboť se jedná o nadindividuální koncept. Koneckonců to ve výše uvedeném úryvku připouští i sám Kubát. Pak je ovšem název publikace zavádějící, přesnější by byla *Kvantitativní analýza žánrů u Karla Čapka* — prezentované metody jsou sice bezesporu aplikovatelné na texty jakéhokoli autora, ovšem jejich výsledky a užitečnost jsou v praxi testované pouze na Čapkovi.

Autorsky rozmanitý korpus je tedy pro obecný popis žánrů nutnou podmínkou. Klíčem k tomu, proč tomu tak je, je právě onen jedinečný autorský styl: za charakteristické pro žánr by mělo být považováno to, co se spolehlivě projeví napříč autory, různým autorským stylům navzdory. Kubát argumentuje, že není možné v tomto scénáři určit, „zda získané hodnoty vypovídají spíš o autorovi, nebo o žánru“. Přitom je to naopak triviální: čím víc autorů se na nějakém rysu shoduje a čím menší je v rámci rysu rozptyl hodnot a jasnější sdílená tendence, tím víc rys vypovídá o daném žánru. Za prvek autorského stylu nelze označit něco, na čem se systematicky shoduje valná většina autorů; stejně tak za rys žánru nelze označit něco, co nacházíme jen u jednoho

z nich. Jinými slovy, jednotlivé autorské styly představují výkyvy různými směry, šum, který se navzájem vykrátí; to, co z něj systematicky vyčnívá, co je sdílené, ukazuje na kolektivní chápání náplně daného žánru. Na tomto přístupu je nakonec přece založená i kvalitativní deskripce: badatel intuitivně hledá prominentní sdílené rysy, jen váhy, které jim přiděluje, jsou subjektivní, tj. ne tak těsně navázané na empirii. Není tedy důvod, proč by podobně nemohla postupovat i deskripce kvantitativní.

Úhrnem, autorský styl a žánr jsou dvě strany téže mince. V obou jde o zapojení různých jazykových prostředků; v prvním případě jde o vybočení z konvence, v druhém o její naplnění, ale fundamentálně sledujeme tytéž jazykové jevy, liší se jen jejich rozložení v populaci. Z diachronního pohledu jsou tyto kategorie dokonce prostupné: uznávaný autor může ovlivnit směřování žánru, když ostatní začnou přejímat prvky jeho autorského stylu. Na prvcích samotných se nic nezmění, ale postupem času se stanou indikátory žánru, ne autora, který je zpopularizoval. Tato proměna vnímání se bude odvíjet od toho, jak se prvky rozšíří v populaci autorů. Abychom mohli zkoumat jak žánr, tak autorský styl, musíme tedy o této populaci mít představu. V kvalitativní studii může být přítomna jen implicitně: píšu-li o Čapkově stylu, tiše se předpokládá, že jsem v životě četl i jiné autory, a mám tedy představu, co činí Čapka Čapkem. Ve studii kvantitativní je potřeba srovnávací data explicitně zahrnout do analýzy — jinak nelze jejich vliv kvantifikovat.

Ve světle této zásadní metodologické výhrady budou pochopitelně jakékoli další připomínky k recenzované publikaci působit druhořadě, nicméně některé aspekty práce s daty je potřeba okomentovat, už jen proto, že datová analýza je ošemetná disciplína a svého druhu umění, takže cit pro ni je dobré si tříbit na konkrétních příkladech. Samotná akvizice dat, tj. naměření hodnot výše jmenovaných indexů pro jednotlivé texty, byla bezpochyby v pořádku — jak již bylo řečeno, v této oblasti se autor pohybuje s velkou jistotou, a navíc v některých případech pro výpočet existuje dedikovaný software. Už méně se ztotožňuji s následným statistickým vyhodnocením, vizualizacemi a interpretací.

Máme-li naměřené hodnoty např. MATTR pro texty různých žánrů, dalším krokem je porovnat skupiny textů podle žánrů a zjistit, do jaké míry se mezi sebou liší. Jednoduchým standardním vizualizačním nástrojem pro tyto účely jsou krabicové grafy, které úsporně znázorňují rozdělení hodnot v rámci žánrů: pokud se krabice dvou žánrů víceméně překrývají, je jasné, že MATTR mezi nimi nerozlišuje. Naopak, čím dál jsou od sebe navzájem, tím jsou oba žánry vzhledem ke zvolenému ukazateli (MATTR) rozdílnější.

Kubát volí jinou cestu, a to cestu testů statistické významnosti rozdílů mezi každým žánrem a všemi ostatními. Už to je problematické, protože pro 8 žánrů je potřeba provést $8 \times 7 \div 2 = 28$ testů, přičemž každý test s sebou nese riziko, že jsem texty do korpusu vybral zrovna tak, že to na základě vybraného vzorku bude vypadat, že se např. pohádka a román z hlediska MATTR statisticky významně liší, ale kdybych měl k dispozici celou populaci, tak by se ukázalo, že mezi nimi rozdíl není. Míru rizika falešně pozitivního výsledku, která je v kontextu výzkumu přijatelná, tzv. hladinu α, si určuje badatel před spočítáním testu; Kubát volí často používaných 0,05 (viz např. s. 49). Problém je samozřejmě v tom, že čím víc takových testů provedu, tím víc narůstá riziko, že některý z nich bude falešně pozitivní.

Testy statistické významnosti jsou stavěné na ověření jedné konkrétní hypotézy, kterou testujeme proto, že pro ni máme nějakou smysluplnou motivaci, např. *pohádky jsou psané pro děti, lze u nich tedy očekávat jednodušší jazyk, a tedy nižší slovní bohatství (MATTR) než u románů*. Jakmile začneme bezhlavě porovnávat napříč všemi možnými kategoriemi, které nám data skýtají, pravděpodobnost statisticky významného výsledku strmě roste — ale s ní i riziko toho, že výsledek bude falešně pozitivní. Provedeme-li 28 testů, každý na hladině $\alpha = 0{,}05$, je celková hladina α výrazně vyšší a u pozitivních výsledků nelze tvrdit, že byla potvrzena statistická významnost na hladině $\alpha = 0{,}05$. Je potřeba počítat s korekcí pro vícenásobné testování, která jednotlivé testy penalizuje, resp. vyžaduje extrémnější dílčí výsledky, tak jako to dělají např. *post-hoc* testy pro dodatečné rozlišení jednotlivých úrovní kategoriální proměnné po analýze rozptylu.

V publikaci lze opakovaně nalézt formulace následujícího charakteru: „Na základě výše uvedených výsledků, *kde více než polovina rozdílů je signifikantních*, můžeme tvrdit, že slovní bohatství je z hlediska diferenciace žánrů poměrně silný nástroj" (s. 50; zvýraznil DL). Zaprvé, počet signifikantních rozdílů je poměrně neotřelý způsob, jak kvantifikovat vztah dvou proměnných. Obvyklejší by bylo spočítat, kolik variability v naměřených hodnotách slovního bohatství lze přičíst na vrub rozdílů v žánrech, tak jako to činí např. výše jmenovaná analýza rozptylu. Zadruhé, i tak nejsou tyto rozdíly signifikantní vzhledem k uvedené hladině α, protože nebyla aplikována žádná korekce pro vícenásobné testování.

Na okraj zmiňme, že u volby samotného statistického testu patrně panuje jisté zmatení. Kubát ho popisuje bez odkazů jako „asymptotický *u*-test, který je ve statistice známý také jako *z*-test" (s. 48). Nejsem si jist, co se zde míní termínem asymptotický; většinou se spíš hovoří o tom, že hodnoty udávané jedním testem se za jistých podmínek asymptoticky blíží hodnotám udávaným testem jiným. V každém případě, Wilcoxonův *u*-test je testem neparametrickým, tzn. robustním vůči datům, která nejsou normálně rozdělená, kdežto *z*-test je parametrický a normální rozdělení předpokládá, a vzorec, který Kubát uvádí na s. 49, neodpovídá ani jednomu z nich, nýbrž tzv. Welchovu *t*-testu,[2] který je také parametrický (pracuje s aritmetickým průměrem a směrodatnou odchylkou). Označovat spočítanou testovou statistiku jako *u* je tedy zavádějící, neboť to navozuje očekávání robustnosti vůči nenormalitě, které tato statistika nesplňuje.

Podobně matoucí jsou bohužel i vizualizace: obecně lze říct, že jich je zbytečně mnoho a ne vždy jsou šťastně zvolené. Např. (ne)závislost zvolených indexů na délce textu dokládá Kubát sedmi bodovými grafy znázorňujícími vztah mezi délkou textu a naměřenou hodnotou indexu (obr. 7–13 na s. 35–38), přitom přesvědčivější, úspornější a přehlednější by bylo uvést korelační koeficienty. Stejně tak vizualizace výsledků klasifikace žánrů pomocí n-gramové analýzy prostřednictvím ne méně než 20 (!) obrázků (obr. 43–62 na s. 95–106) svědčí o jisté bezradnosti. V průběhu výzkumu přirozeně dává smysl vytvořit a prozkoumat větší množství grafů, úkolem badatele ale při psaní odborného textu je vybrat ty, které jsou relevantní pro zodpovězení zvolené výzkumné otázky, a úzce je propojit s textem. Neměl by čtenáři předložit velké

2 Za dohledání testu, kterému vzorec ve skutečnosti odpovídá, vděčím Václavu Cvrčkovi.

množství grafů, které s textem souvisejí jen volně, a nechat jej, ať v nich smysl hledá sám. U obr. 63, který představuje nečitelný dendrogram všech 760 analyzovaných textů, sám Kubát uvádí, že „vzhledem k velkému množství textů graf slouží spíš jen ilustrativně" (s. 108), a připojuje odkaz ke stažení verze v barvě a vyšším rozlišení. Obsah obrázku přitom nijak nekomentuje, není tak jasné, co si z něj čtenář má odnést. Takový obrázek je zbytečný, pouze odvádí pozornost, nemluvě o tom, že kdyby byl napojen na nějakou interpretaci, bylo by možné z něj vydestilovat důležité informace a upravit ho tak, aby byl srozumitelný i v tištěné podobě.

Jeden typ grafu pak zřejmě představuje autorovu inovaci: pro každý stylometrický index se snaží zobrazit vzdálenosti mezi žánry pomocí dvourozměrného grafu (např. obr. 18 pro MATTR). V případě jediného indexu jsou ovšem pochopitelně i vzdálenosti jednorozměrné, stačilo by tedy vynést body odpovídající jednotlivým žánrům na přímku. Kubát místo toho doplňuje druhou osu, která znázorňuje sumu testových kritérií *u* z testů, jichž se žánr účastnil. Vzhledem k výše rozebírané definici tohoto kritéria jde v podstatě o normalizovanou sumu vzdáleností žánru od ostatních žánrů, tj. jiný pohled na tytéž hodnoty, které jsou již zobrazeny na první ose. Rozklíčovat takovou vizualizaci není snadné, protože je zbytečně složitější než data, z nichž vychází. Přitom smyslem vizualizace by naopak mělo být složitější data zjednodušit a zpřístupnit, ne znepřehlednit.

Ve shrnující kap. 4.8 pak již nepřekvapí, že Kubát porovnává užitečnost jednotlivých stylometrických indexů pro rozlišování žánrů také podle dosažené celkové sumy testových kritérií *u*, resp. podle počtu statisticky významných rozdílů. Jak jsme nastínili výše, agregace dílčích statistických testů není šťastný nápad, standardní postup by byl zvolit statistickou metodu, která dokáže vzít v potaz všechny cílové proměnné naráz. V závěru se Kubát vrací k problematice zobecnitelnosti výsledků výzkumu: „získané výsledky a dílčí závěry bude nutné podpořit v dalších výzkumech zejména (a) rozšířením analýz o další autory, žánry a jazyky, (b) použitím dalších stylometrických indexů a metod" (s. 121). Z mého pohledu je doplnění dalších žánrů nutnost a mělo být součástí již této publikace, aby dostála svému názvu. Nevím, zda si to autor představoval tak, že vzniknou metodologicky obdobné výzkumy dalších autorů, které se budou následně propojovat přes součty *u* a počty statisticky významných rozdílů. Pokud ano, pak doufám, že je z výše uvedeného dostatečně jasné, že toto není cesta kupředu. Pokud ovšem bude Kubát napříště věnovat statistickému vyhodnocení a interpretaci stejnou péči jako rešerši a aplikaci stylometrických ukazatelů, zní tento program slibně.

LITERATURA:

Cvrček, V., & Chlumská, L. (2015). Simplification in translated Czech: A new approach to type-token ratio. *Russian Linguistics*, *39*(3), 309–325.

David Lukeš | Ústav českého národního korpusu FF UK
<david.lukes@ff.cuni.cz>