Nesamozřejmá otevřenost lingvistického výzkumu

Non-self-evident openness of linguistic research

Václav CVRČEK | Ústav Českého národního korpusu FF UK
Jan CHROMÝ | Ústav českého jazyka a teorie komunikace FF UK

1 Úvodem

Cílem našeho příspěvku Lingvistika jako otevřená a transparentní disciplína (s. 5–14) bylo iniciovat v českém prostředí diskusi o potřebě sdílení dat a dalších materiálů spojených s realizací výzkumu. Nastínili jsme, co toto sdílení představuje, v čem je podle našeho názoru potřebné a výhodné, a představili jsme několik příkladů dobré praxe, které tyto výhody ilustrují. Na závěr textu jsme uvedli několik návrhů pro české akademické časopisy a instituce.

V reakci na náš text vzniklo jedenáct odpovědí. Všem jejich autorům velmi děkujeme. Ať už se jednalo o příspěvky souhlasné, či kritické, ukazuje tato diskuse, že otevřenost a transparentnost lingvistiky je téma, jehož důležitost je pociťována badateli s různými teoretickými východisky a odbornými zaměřeními.

Velmi děkujeme za všechna doplnění našich výchozích myšlenek. Radek Čech (s. 15–16) a Petr Pořízka (s. 22–27) například poukazují nejen na samotné sdílení dat a výzkumných materiálů, ale i na dostupnost výsledných studií, což je pochopitelně zcela zásadní. Souhlasíme s tím, že by české akademické časopisy měly (i vzhledem k minimálním výnosům z prodeje) fungovat v režimu open access a že placený přístup k textům efektivně blokuje šíření lingvistických poznatků. Silvii Cinkové (s. 16–17) vděčíme za doplnění o klíčový právní koncept Fair Use, Mojmír Dočekal (s. 19–20) připomíná široce využívanou platformu GitHub, Jaroslav David dodává hned několik příkladů dobré praxe, Daniel Vrbík a Václav Lábus (s. 20–22) podrobně informují o zajímavém projektu Živá jména a Petr Pořízka (s. 22–27) představuje Olomoucký mluvený korpus a další otevřené projekty.

V reakcích zazněly rovněž kritické či nesouhlasné komentáře. Právě jim chceme v tomto závěrečném slově věnovat pozornost. Domníváme se, že nesouhlasná vyjádření vyplývají z části z určitého nedorozumění a že se v posledku s autory těchto vyjádření shodneme. V jiných případech však zastáváme skutečně odlišný názor a pokusíme se podrobněji vysvětlit proč. Naše reakce jsme se rozhodli strukturovat do šesti tematických bloků.

2 Sdílení dat a dalších materiálů není spojeno jen s kvantitativním přístupem k jazyku a jeho užívání

Myšlenka, že se naše názory na transparentnost a otevřenost lingvistiky pojí úzce s kvantitativním, či dokonce experimentálním přístupem k jazyku a jeho užívání, se objevuje hned v několika reakcích na náš původní text (Pytlíková – Černá – Šimek, s. 28–29; Štěpán – Wojnarová, s. 31–33; Kaderka – Sherman – Havlík, s. 33–38; Stupňánek, s. 38–42). Domníváme se, že se patrně jedná o nedorozumění, které je dáno tím, že některé z argumentů pro zveřejňování a sdílení dat a dalších materiálů dávají smysl primárně v kvantitativním rámci. Z našeho hlediska je však otevřenost a transparentnost lingvistiky v tomto smyslu stejně podstatná jak pro kvalitativní, tak pro kvantitativní přístupy.

Z devíti argumentů, které jsme uvedli ve prospěch sdílení dat a dalších materiálů, se kvalitativního výzkumu stejně dobře jako kvantitativního týká přinejmenším sedm: (1) zpětná ověřitelnost, (2) prevence, (4) vzájemná výpomoc a otevírání nových možností, (5) vzdělávání studentů,

(6) společenská a morální zodpovědnost, (7) snadná dohledatelnost a zpětná interpretovatelnost, (8) zveřejnění dat jako výstup (s. 8–11). Jinými slovy, jediné dva argumenty, o jejichž smysluplnosti lze z perspektivy kvalitativních přístupů pochybovat, jsou (3) replikovatelnost a (9) metaanalýza.

Zde je dobré podotknout, že replikovatelnost není spjata s kvantitativním přístupem jako takovým (jak naznačují Kaderka, Sherman a Havlík, s. 33–38), ale spíše s jevy, u nichž předpokládáme určitou stabilitu v čase. Například kvantitativní studii Pavla Jančáka (1974) o hlavních hláskoslovných rysech v mluvě pražské mládeže replikovat v pravém slova smyslu nemůžeme, protože prostě nemáme k dispozici pražskou mládež ze 70. let 20. století a Praha se navíc zásadně změnila z demografického hlediska. Stejně tak jsme si plně vědomi toho, že nejde replikovat kvalitativní výzkum interakcí v houslařské dílně, který popisují Kaderka, Sherman a Havlík (s. 35), protože se tyto interakce neopakují, jsou časoprostorově jedinečné. V takových případech je však zásadní **reprodukovatelnost** výzkumu (viz např. Berez-Kroeker et al., 2018). Z hlediska vývoje vědy považujeme za zcela zásadní, abychom byli schopni každý výzkum udělat s časovým odstupem co možná nejvíc stejně. To nám totiž umožňuje srovnání v čase, respektive třeba napříč různými typy situací apod.

Reprodukovatelnost je důležitá i proto, že nám umožňuje výzkum modifikovat na základě zkušenosti z předchozích pokusů a poučit se z problémů se způsobem jeho realizace. V tomto ohledu se domníváme, že jsme s reakcí Kaderky, Sherman a Havlíka (s. 33–38) zcela v souladu. Zdůraznili bychom zde to, že jsme v našem výchozím textu nemluvili pouze o sdílení dat, ale i o sdílení dalších materiálů. Ty mohou být velmi různorodé a s ohledem na zaměření výzkumu i velmi specifické.

Zveřejnění dat je nicméně prospěšné i v případě, kdy výzkum nelze z praktických důvodů přesně replikovat (viz oba výzkumy zmiňované výše). V případě, kdy jsou nasbíraná data k dispozici, lze zpětně ověřit, zda závěry odpovídají pozorování (samozřejmě s nevyhnutelnou mírou zkreslení).

Potřebu replikovatelnosti kritizuje i Stupňánek (s. 38–42) a příčiny problémů, které má replikovatelnost odstraňovat, spatřuje téměř výlučně v testování statistické významnosti nulové hypotézy (*null hypothesis significance testing*; NHST). Problémy související s NHST nijak nepodceňujeme a pozitivní vliv replikací na potlačení některých jejích nezdravých projevů nezpochybňujeme, otázku transparentnosti však chápeme výrazně šířeji a přímou souvislost NHST s replikovatelností nevidíme (potřeba replikovatelnosti je zcela zásadní třeba i u výzkumů využívajících bayesovskou statistiku, tedy nikoliv NHST).

3 Sdílení dat a dalších materiálů je podřízeno etickým standardům

Několik reakcí (Kapišovská et al., s. 29–31; Štěpán – Wojnarová, s. 31–33; Kaderka – Sherman – Havlík, s. 33–38) doplňuje pohled na celou problematiku o etické aspekty spojené se sdílením dat a dalších materiálů. Ztotožňujeme se s názorem, že ne vše je možné z etických či právních důvodů zveřejňovat. Běžnou praxí je anonymizace dat před jejich zveřejněním, aby nebylo dohledatelné, od koho pochází (zcela specifický je v tom výzkum onomastický, jak upozorňují Štěpán a Wojnarová (s. 31–33), kde anonymizaci z podstaty zkoumaného předmětu provést nelze).

Uznáváme, že v našem výchozím příspěvku nebyla tato nadřazenost etických standardů vůči potřebě sdílení dat dosti zdůrazněna, z čehož patrně vznikla jistá nedorozumění. Obecně se domníváme, že zveřejňovat by se mělo ideálně maximum možného, a to z již uvedených důvodů. Tam, kde zveřejnění brání právní či etické důvody, je třeba postupovat individuálně; klíčovou roli by v takových případech měla hrát komise pro etiku výzkumu (ať už v rámci výzkumných organizací, nebo při redakčních radách, hodnotících panelech apod.). To souvisí i s naším návrhem, aby zveřejnění dat po autorech požadovaly akademické časopisy. Tento návrh nemyslíme kategoricky,

ale opět se zřetelem k etickým standardům. Je zcela v pořádku, není-li možné některá data zveřejnit, nebo není-li je možné zveřejnit v plném rozsahu. Ale správné podle nás není data využívaná v publikované studii nezveřejnit, pokud to možné je.

4 Lingvistika není výjimkou ve sdílení dat a dalších materiálů

Dočekal (s. 19–20) a Pořízka (s. 22–27) poukazují na to, že téma otevřenosti a transparentnosti není zdaleka jen lingvistickou otázkou. S tím pochopitelně zcela souhlasíme. Naopak z odpovědi B. Stupňánka (s. 38–42) se zdá, jako by lingvistika měla být vůči těmto problémům imunní. Stupňánek nám například vyčítá, že neuvádíme žádný lingvistický příklad, za který patrně nepovažuje příklad neúspěšné replikace výzkumu Lery Boroditské (2001) pro jeho přílišnou psychologičnost. Máme-li tedy zůstat v úzkém chápání výzkumu jazyka a jeho užívání, můžeme doporučit například studii Edwarda Gibsona a Eveliny Fedorenkové (2013) The need for quantitative methods in syntax and semantics research, v níž jsou rozebírány tři ilustrativní příklady z klasických lingvistických domén, tedy z výzkumu syntaxe a sémantiky.

5 Sdílení dat a dalších materiálů má být výhodné

V několika reakcích (David, s. 17–18; Pytlíková – Černá – Šimek, s. 28–29; Stupňánek, s. 38–42) se objevil důraz na aktuální systém financování vědy, který sdílení dat nepodporuje. Ačkoli tato problematika dalece přesahuje cíl úvodního příspěvku a v zásadě i možnosti jakékoli vědecké diskuse, souhlasíme s tím, že by situace měla být v tomto ohledu lepší, než je dnes.

Nedomníváme se však, že by za současného stavu představovalo sdílení dat a dalších materiálů nějakou nevýhodu. Už několik let je v Rejstříku informací o výsledcích (RIV) položka „specializovaná veřejná databáze“, která umožňuje sdílená data vykázat. Kromě toho lze zveřejnění dat uvést jako regulérní výstup většiny typů vědeckých grantových projektů či v rámci kvalifikačních či kariérních procesů. To však není klíčové. Jsme přesvědčeni, že samotné zveřejnění dat může mít v řadě případů mnohem vyšší dopad na danou disciplínu (respektive určitý její výsek) než třeba studie či kniha, která je na těchto datech založena. Důvodem je především to, že se sběrem a zpracováním dat je spojeno netriviální množství expertní práce, což jejich hodnotu pro další badatele významně zvyšuje. Vedle bibliometrického „výnosu“ je tedy vhodné zvažovat i obecný „přínos“ sdílených dat.

Za zcela odstrašující považujeme Stupňánkovu (s. 40) obhajobu nesdílení dat tím, že „chovat se co nejsobečtěji“ je „nejvýhodnější strategie“. I kdybychom přistoupili na to, že takovýto postup je tou nejvýhodnější strategií k získání finančních prostředků v aktuálních podmínkách hodnocení vědy v ČR, pevně doufáme, že se výzkum ve většině pracovišť (včetně dialektologického oddělení ÚJČ AV ČR, v němž Stupňánek pracuje) řídí spíše snahou o zmnožení vědeckého poznání toho, jak funguje jazyk. Na rozdíl od Bronislava Stupňánka vidíme klíč k posunu v pochopení toho, jak svět kolem nás funguje, ve spolupráci, nikoli v izolaci.

6 Jediný smysl dat je v jejich použití

V několika reakcích byla zmíněna obava, že by sdílení dat mohlo vést k tomu, že je někdo vytěží místo nás. Jaroslav David (s. 17–18) to ilustruje na rozdílu mezi lovci a sběrači. Jakkoliv je tato metafora působivá, nemyslíme si, že by vystihovala situaci přesně. V prvé řadě každý, kdo sbírá nějaká data, může při tom využívat dosud zveřejněná data a materiály (třeba strukturaci rozhovoru, standardizované dotazníky, způsoby přepisu apod.), může postupovat stejně jako nějaký předchozí výzkum, anebo může postup využitý v předchozím výzkumu záměrně modifikovat tak, aby se

vyhnul problémům, na které předchozí výzkum narážel.[1] Zároveň „sběrač“ má u vlastních dat před „lovcem“ velkou výhodu – jednak na svých datech může pracovat poměrně dlouhou dobu předtím, než jsou zveřejněna (ostatně, nejběžnější způsob zveřejnění dat a dalších materiálů je v návaznosti na publikaci studie, která z těchto dat vychází), jednak tato svá data dobře zná. Je vysoce nepravděpodobné, že by k těmto datům přišel nějaký „lovec“ a podnikl na nich právě tu analýzu, již plánoval provést „sběrač“. Zkrátka a dobře, ryzí „lovecká“ strategie by byla velmi nevýhodná – spoléhat na to, že data, která budou teprve zveřejněna, bude možné použít k vlastnímu výzkumu, by bylo krajně rizikové a lovcovy výzkumné možnosti by se musely pružně podřizovat tomu, co kdo zrovna zveřejní. To přitom nikterak neumenšuje hlavní poselství naší výzvy, že sdílení dat přináší potenciální užitek – právě naopak. Pouze poukazujeme na to, že „sběračům“ nevzniká zveřejněním jejich dat a dalších materiálů žádná podstatná škoda.

Neztotožňujeme se s argumentem, který do diskuse o prospěchu sdílení dat přináší Stupňánek (s. 40), když tvrdí, že „prospěch z dostupnosti velkého množství dat budou mít obzvláště kvantitativní obory lingvistiky“. Představme si, že bychom měli v online databázi k dispozici audio- či videozáznamy nejrůznějších nářečních projevů s jejich celými přepisy, které by byly snadno prohledatelné, třeba i díky tomu, že by byly lemmatizované a jinak anotované. Nebyl by to přínos primárně pro dialektologii jako takovou? Příklad ze Slovenska, kde byl ve spolupráci se SNK zveřejněn Korpus nárečí (viz KNSNK), to soudě podle pozitivních ohlasů tamních dialektologů poměrně zřetelně dokládá.

7 Náročné je zpracování dat, nikoliv jejich sdílení

Několik reakcí poukazuje na náročnost zpracování dat (viz Kapišovská et al., s. 29–31; Pytlíková – Černá – Šimek, s. 28–29; Stupňánek, s. 38–42) a dlouze se jí pak věnují Kaderka, Sherman a Havlík (s. 33–38). Souhlasíme s tím, že se jedná o důležitou otázku. Na druhou stranu se nedomníváme, že by zveřejňování samo o sobě mělo zásadní vliv na náročnost zpracování dat. Nepromyšleným nakládáním s daty dochází k jejich ztrátám, duplicitnosti, respektive existenci více různých verzí téhož, nejasné zpětné kontrole nad různými rozhodnutími při zpracování dat či jejich získávání apod., což však vůbec nesouvisí s jejich zveřejňováním.

Uvedené důvody vedou badatele nově k tomu, že si ještě před započetím výzkumu vytváří plán managementu dat, který se v průběhu výzkumu postupně aktualizuje. Součástí plánu jsou mimo jiné informace o sběru dat, jejich ukládání a dokumentaci, o způsobu ukládání a zálohování, ale třeba i o zodpovědnosti jednotlivých členů týmu apod. Plány managementu dat představují práci navíc, na kterou badatelé nebyli dříve zvyklí. Na druhou stranu však samotným badatelům v posledku významně pomáhají a zkvalitňují celý výzkumný proces. Na vedení takových plánů dnes existuje například portál DMP Online (viz DMP), jehož použití je bezplatné. Zároveň se zdá, že do budoucna bude plán managementu dat vyžadován grantovými agenturami jako nutná součást samotných žádostí. Pokud si badatel vede plán managementu dat, postupuje zodpovědně a vyvaruje se

[1] Je nesporné, že v takových situacích se sluší na inspirační zdroj odkázat. To je i případ Olomouckého mluveného korpusu (OMK), o němž píše ve svém příspěvku jeho autor Petr Pořízka (s. 22–27) ve vztahu k druhé generaci mluvených korpusů ČNK. Na doplnění toho, jak celou věc líčí Petr Pořízka, bychom rádi uvedli, že koncepce OMK nebyla „převzata“, byl pouze jedním z několika inspiračních zdrojů pro korpusy ORAL a ORTOFON (v mnoha ohledech se s OMK rozcházejí, navazují na jiné projekty nebo jdou svou vlastní cestou). Za klíčové v této souvislosti dále považujeme to, že v textech, kde byla koncepce mluvených korpusů ČNK představována, je OMK jako inspirační zdroj řádně citován.

různých problémů, které ho mohou při realizaci výzkumu potkat. Zároveň je pro něj pak snadné cokoliv zveřejnit, protože data má plně pod kontrolou, má k nim potřebnou dokumentaci apod.

Sdílení dat a dalších materiálů zároveň souvisí s tím, co jsme v našem výchozím textu nazvali „fundamentální změnou celého přístupu k vědě a výzkumu, která má zásadní dopady na výzkumnou praxi“ (s. 6). Zpracování dat do podoby, v které je lze zveřejnit, je spojeno s určitým úsilím, ale tuto práci „navíc“ vnímáme jako součást přirozeného vývoje disciplíny, a věříme proto, že se v zájmu snahy o zvyšování úrovně vědecké produkce stane inherentní součástí naší vědecké rutiny, stejně jako třeba korektura článku nebo kontrola bibliografie.

8 Závěrem

Na závěr bychom chtěli ještě jednou poděkovat všem diskutujícím, jakož i redakci Naší řeči, která poskytla diskusi o otevřenosti a transparentnosti výzkumu jazyka a jeho užívání prostor. Věříme, že nás všechny tato diskuse posune dál a budeme více přemýšlet o tom, jakým způsobem s daty vlastně nakládáme a jak bychom s nimi mohli nakládat tak, aby to bylo prospěšnější vědě samotné. Zlepšení nepřijde automaticky, a rozhodně ne, pokud budou principy obhajované v úvodním textu naplňovány pouze formálně. Sdílení dat a dalších materiálů je nicméně podle nás výrazným krokem na cestě k vyšší transparentnosti výzkumu. Ne jediným, ale výrazným.

LITERATURA

Berez-Kroeker, Andrea L. – Gawne, Lauren – Kung, Susan Smythe – Kelly, Barbara F. – Heston, Tyler – Holton, Gary – Pulsifer, Peter – Beaver, David I. – Chelliah, Shobhana – Dubinsky, Stanley – Meier, Richard P. – Thieberger, Nick – Rice, Keren – Woodbury, Anthony C. (2018): Reproducible research in linguistics: a position statement on data citation and attribution in our field. *Linguistics*, 56(1), s. 1–18.

Boroditsky, Lera (2001): Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive psychology*, 43(1), s. 1–22.

Gibson, Edward – Fedorenko, Evelina (2013): The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1–2), s. 88–124.

Jančák, Pavel (1974): Frekvence hlavních hláskoslovných znaků v mluvě pražské mládeže. *Naše řeč*, 57(4), 191–200.

KNSNK: *Korpus nárečí Slovenského národného korpusu – dialekt-4.0* [online] (2018). Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV. Cit. 16. 12. 2020. <https://korpus.juls.savba.sk>.

DMP: *DMPonline: Data Management Plans that Meet Institutional Funder Requirements* [online]. Cit. 16. 12. 2020. <https://dmponline.dcc.ac.uk/>.

Václav Cvrček
Ústav Českého národního korpusu FF UK
Panská 890/7, 110 00 Praha 1
vaclav.cvrcek@ff.cuni.cz

Jan Chromý
Ústav českého jazyka a teorie komunikace FF UK
náměstí Jana Palacha 2, 116 38 Praha 1
jan.chromy@ff.cuni.cz