Lingvistika jako otevřená a transparentní disciplína[1]

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

Linguistics as an open and transparent science

The aim of this paper is to initiate the discussion on current trends in sharing research data (and other materials related to the research process) in the Czech linguistic community. First, we present a brief explanation of what the primary incentives are for data sharing. Second, we discuss nine main reasons for sharing data and other research materials. These are verifiability of the results, prophylactic effects, replicability, synergy and cooperation between the researchers, use of data in teaching, social and ethical responsibility, easy accessibility and interpretability of the results, data as an outcome, and the possibility to conduct a meta-analysis. Third, we focus on several examples of good practice, i.e. the Czech National Corpus, the LINDAT/CLARIAH-CZ infrastructure, TROLLing, and the Open Science Framework. As a conclusion of the paper, we present three appeals for Czech linguistics: (a) Academic journals should require data sharing as a standard aspect of the publication process, (b) academic institutions should require sharing of data and other research materials by the academic staff, (c) universities should adopt clear rules for sharing of data and other research materials for students.

Key words: data sharing, open data, replicability, reproducibility, transparency, verifiability

Klíčová slova: otevřená data, ověřitelnost, replikovatelnost, reprodukovatelnost, sdílení dat, transparentnost

Cílem tohoto textu je zahájit v české lingvistické komunitě diskusi o aktuálních trendech ve světovém výzkumu, které se týkají otevřenosti ve sdílení dat a dalších materiálů spojených s realizací výzkumu.[2] Ačkoliv nejsou pro české prostředí tyto otázky zcela nové (a jak si ukážeme, existuje hned několik příkladů dobré praxe), má v tomto ohledu česká lingvistika podle našeho soudu dosud značné rezervy.

Zahájení diskuse o těchto záležitostech považujeme za zásadní nikoli proto, že se jedná o aktuální trend ve světové vědě. Celou otázku vnímáme v kontextu dlouhodobé snahy o větší dostupnost vědy (nikoli pouze lingvistiky), která se prosazuje

[1] Tento článek vznikl v rámci projektu GA ČR *Lingvistické faktory pochopitelnosti v českých administrativních a odborných textech* (19-19191S) a v rámci projektu *Český národní korpus* (LM2018137) financovaného Ministerstvem školství, mládeže a tělovýchovy v rámci aktivity Projekty velkých infrastruktur pro VaVaI. Práce na tomto textu byla podpořena také Ministerstvem školství, mládeže a tělovýchovy České republiky v rámci projektu *LINDAT/CLARIAH-CZ* (LM2018101).

[2] Za cenné připomínky k první verzi tohoto textu by autoři rádi poděkovali E. Lehečkové a M. Křenovi.

zejména konceptem open access[3] (jenž si nejvíc spojujeme s volnou dostupností odborných textů). Na tento trend navazuje a rozvíjí ho obecnější koncept open science, který požaduje svobodné sdílení také výchozích dat, metod, zdrojových kódů atp. Rádi bychom přitom hned v úvodu zdůraznili, že se nejedná pouze o otázku technickou, ale o fundamentální změnu celého přístupu k vědě a výzkumu, která má zásadní dopady na výzkumnou praxi.[4]

Co znamená sdílení dat a dalších materiálů?

Lingvistika na celém světě prošla v posledních desetiletích silnou empirizací. Některé dílčí lingvistické či pomezní disciplíny (jako třeba sociolingvistika a psycholingvistika) jsou empirické ze své podstaty, jiné se tímto směrem vyprofilovaly postupně (například výzkum gramatiky), nicméně prakticky pro všechny oblasti zkoumání jazyka a jeho užívání v dnešní době platí, že se v nich pracuje s daty (byť taková data mohou být nejrůznějšího typu – například soubory textů, nahrávky a jejich přepisy, reakční časy na určité podněty, odpovědi na otázky v dotaznících, excerpované jazykové jednotky apod.). Právě na analýze dat jsou založeny následné závěry a zobecnění, o kterých se mohou čtenáři dočíst v různých studiích, knihách či studentských závěrečných pracích. Práce s daty je tedy něco, co lze v dnešní době chápat jako integrální součást lingvistického provozu.

Moderní počítačová technologie nás posouvá nejen ve způsobech zpracování dat, ale i v možnostech jejich sdílení s ostatními badateli prostřednictvím specializovaných digitálních platforem. Aby sdílení plnilo svůj účel, musí data splňovat určité náležitosti, které se někdy shrnují pod akronym FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*), přičemž za klíčové bychom v kontextu české lingvistiky mohli považovat tři aspekty: čitelnost dat, jejich interpretovatelnost a přístupnost.

Aniž bychom chtěli zabíhat do technických detailů, jakkoli jsou s otázkou kurátorství dat nevyhnutelně spojené,[5] je třeba v prvé řadě zmínit to, aby se data zveřejňovala ve **strojově čitelném formátu**, a to s výhledem na to, aby bylo možné s takovými daty pracovat třeba i za několik desítek let, kdy lze předpokládat zásadně odlišné technologie, a to bez nutnosti pořizovat specifický software na jejich čtení.

---

3 Přístupnou formou přibližuje zásady open access pro české prostředí např. portál <https://openaccess.cz>.

4 Náš text se zaměřuje na věcné a metodologické stránky problému, je třeba ovšem dodat, že problematika open access a open science je řešena příslušnými národními i nadnárodními koncepcemi vědecké politiky, naposledy Berlínskou deklarací z roku 2003, k níž se připojily mnohé akademické instituce z ČR. Incentivy pro sdílení dat poskytují i prestižní grantová schémata, např. Horizon 2020 a jeho nástupce Horizon Europe požadují po badatelích publikace „podle zásad paradigmatu otevřeného přístupu".

5 Úvod do tzv. *data curration* pro humanitní disciplíny poskytuje např. kolektivní publikace Digital Humanities Data Curation (viz DHDC).

Preferují se proto zavedené a otevřené standardy, které disponují dobrou dokumentací, např. v případě textových dat může jít o některý z doporučovaných značkovacích jazyků (XML), v mnoha případech ale postačuje prostý text bez formátování.

Aby byla data náležitě interpretovatelná, je klíčové, aby byla zveřejňována spolu s podrobnou dokumentací, tedy jakýmsi průvodním textem, který popisuje strukturu a celkovou povahu datového souboru, ideálně včetně pravidel kódování apod. Samotná data totiž mohou být poměrně komplikovaná a pro někoho mimo původní badatelský tým i těžko uchopitelná (používají se například různé zkratky, nemusí být jasné, o jaké proměnné se jedná apod.).

Sdílení dat pochopitelně automaticky neznamená, že jsou data přístupná komukoliv a že si s nimi kdokoliv může dělat, co ho napadne. Sdílení podléhá určitým právním omezením a typicky je spojeno s různými typy licencí. Data mohou být dostupná volně, anebo přístup k nim může být určitým způsobem omezený (a případný zájemce o tato data musí podepsat ujednání o tom, pro jaké účely smí data využít, že je nesmí šířit apod.). Licencí, pod nimiž lze data sdílet, je k dispozici velké množství, a pro většinu situací tak stačí zvolit nejvhodnější z již existujících (namísto vytváření vlastní licence). Pro sdílení vědeckých dat je pravděpodobně nejpoužívanější soubor mezinárodně uznávaných licencí Creative Commons (CC),[6] který umožňuje omezit redistribuci či komerční využití apod.

Kromě samotných dat je možné sdílet i další materiály. Pokud jsou data badatelem či badatelským týmem zpracovávána kvantitativně, je vhodné v souladu s požadavky open science sdílet rovněž skripty, jejichž prostřednictvím byla analýza prováděna. Touto cestou jde i rozsáhlá evropská iniciativa European Open Science Cloud (viz EOSC), jejímž cílem je vyvinout a provozovat virtuální platformu nejenom pro sdílení dat, ale také pro jejich analýzu pomocí softwarových nástrojů.

V experimentálním výzkumu se pak prosazuje tendence zveřejňovat celý proces výzkumného postupu. První krokem je v takovém případě preregistrace, při které se specifikují všechny plánované metodologické aspekty výzkumu (hypotézy, plánovaný vzorek, technika sběru dat, plánované způsoby statistické analýzy atd.). O důležitosti preregistrací viz například u Noska et al. (2019). V dalším kroku se pak spolu se získanými daty zveřejňují například použité stimuly, skript samotného experimentu (je-li využíván nějaký experimentální software), hypotézy apod. Podstatné přitom je, že jsou stále dostupnější internetové služby, jako je např. portál Open Science Framework (viz OSF; viz též níže), které jsou uživatelsky velmi přívětivé a které umožňují snadno a zdarma sdílet výzkumná data bez nutnosti osvojovat si komplikovaná technická řešení.

[6] České překlady licencí a jejich vysvětlení jsou k dispozici v rámci dedikovaného portálu Creative Commons Česká republika (viz CC ČR).

Proč sdílet data?

Na začátku článku jsme uvedli, že sdílení dat a dalších materiálů je spojeno s fundamentální změnou vědeckého přístupu, kterou lze ve shodě s principy open science popsat jako posun od prostého publikování výsledků či závěrů ke sdílení nejen znalostí, ale i dat pro jejich snadné opětovné využití v navazujícím výzkumu. Důvody, proč sdílet data, tak zdaleka nejsou jen technické, ale spočívají v odlišném chápání toho, jak má být věda provozována. Níže uvádíme devět zásadních argumentů, proč bychom měli data (a další materiály týkající se realizovaného výzkumu) sdílet a jak z toho může celá badatelská komunita profitovat.

1. **Zpětná ověřitelnost**. Sdílení dat a dalších materiálů je nezbytné pro zpětnou ověřitelnost výzkumných závěrů. Při zpracování dat může dojít k řadě nezáměrných pochybení, např. k různým chybám ve výpočtech, nepřesné anotaci, omylům při přepisu apod. Jsou-li data, s nimiž se ve výzkumu pracovalo, vědecké komunitě k dispozici, lze zpětně tyto chyby odhalit, a upřesnit tak závěry, k nimž autoři došli. To je přitom zcela zásadní pro úspěšný rozvoj jakékoli disciplíny, protože nepřesné závěry další výzkum brzdí, nebo ho dokonce svádí na scestí, nemluvě o důsledcích, které můžou chybné závěry mít pro aplikační sféru. Případem tohoto typu mimo lingvistiku, který zároveň dokládá, že podobná pochybení se nevyhýbají ani významným postavám oboru, je kauza výzkumu dvou respektovaných ekonomů z Harvardovy univerzity C. Reinhartové a K. Rogoffa, kteří se v roce 2010 zaměřili na vztah prosperity a zadluženosti státu. Srovnáním údajů z různých zemí dospěli k závěru, že přesáhne-li zadlužení státu vzhledem k jeho HDP hranici 90 %, projeví se to poklesem hrubého produktu v průměru o 0,1 %. Jejich vstupní data v roce 2013 přezkoumal Ph.D. student T. Herndon, prof. M. Ash a prof. R. Pollin (z Univerzity v Massachusetts Amherst), kteří došli k závěru, že vlivem omylu při manipulaci s daty v tabulce je celkový závěr nesprávný – namísto poklesu o 0,1 % z dat plyne růst o 2,2 %.[7]

2. **Prevence**. Z výše uvedeného plyne, že sdílení dat může mít nezanedbatelný profylaktický efekt. Možnost identifikovat chyby v předchozích výzkumech vede samotné badatele k větší pozornosti a metodologické důslednosti. S tímto bodem úzce souvisí preregistrování experimentů, což je dnes již v řadě vysoce hodnocených mezinárodních časopisů vyžadováno (bez preregistrace není možné publikovat studii o daném experimentu). Jestliže autor musí zveřejnit detailní plán svého výzkumu ještě před jeho zahájením, je nucen jednotlivé kroky velmi dobře promyslet a odůvodnit. Účinně se tak brání mimo jiné takzvanému „rybaření“ (*fishing*), tedy nezdravé praxi, kdy badatel nasbírá data a následně v nich „loví“ jakékoliv

[7] Stručné shrnutí celé kauzy spolu s odkazy na původní analýzu a její revizi obsahuje článek dostupný na adrese <https://theconversation.com/the-reinhart-rogoff-error-or-how-not-to-excel-at-economics-13646>.

statisticky významné jevy, bez ohledu na to, zda to nějak souvisí s jeho předchozími hypotézami či původním záměrem, s nímž byla data shromažďována.

3. **Replikovatelnost**. Pokud autoři sdílí data a další materiály, umožňují ostatním výzkum replikovat, respektive na něj účinně navazovat a precizovat dosažené výsledky. V některých oborech (zejména v sociální psychologii) se dnes mluví o replikační krizi – badatelům se často nedaří úspěšně replikovat ani velmi známé výzkumy, o nichž se píše v učebnicích a které představují jistý myšlenkový kánon (viz například Shrout – Rodgers, 2018; Wiggins – Christopherson, 2019). Problém s replikovatelností výzkumu se však nevyhýbá ani disciplínám, které se zabývají výzkumem jazyka a jeho užívání, a to ani velmi známým a široce citovaným výzkumům. Například Lera Boroditsky (2001) publikovala známou studii, v níž experimentálně srovnávala mluvčí angličtiny a čínštiny a došla k závěru, že rodilí mluvčí těchto dvou jazyků přemýšlí o čase odlišně (mluvčí angličtiny horizontálně, mluvčí čínštiny vertikálně). V roce 2007 však vyšly dvě studie, které se neúspěšně snažily výzkum Boroditské replikovat, a vážně tak zpochybnily závěry široce známého a uznávaného výzkumu: David January a Edward Kako (2007) se neúspěšně snažili replikovat výsledky pro rodilé mluvčí angličtiny (replikace se nezdařila ani v jednom ze šesti pokusů) a Jenn-Yeu Chenová (2007) se pokoušela replikovat výsledky pro rodilé mluvčí čínštiny (provedla dvě neúspěšné replikace).

4. **Vzájemná výpomoc a otevírání nových možností**. Sdílíme-li data a další materiály, umožňujeme ostatním, aby s nimi pracovali v analýzách, které sami neplánujeme. Samotný sběr dat je obvykle poměrně náročná část výzkumného projektu, která vyžaduje hodně času a často i finančních prostředků. Je proto škoda, je-li vynaložená aktivita využita jen omezeně. Zkusme se zamyslet nad tím, kolik dat bylo jenom v posledních deseti letech nasbíráno v různých studentských závěrečných pracích na jednotlivých bohemistických pracovištích v České republice a co se s těmito daty stalo. Kdyby byla jenom tato data šířeji dostupná (a kvalitně anotovaná), umožnilo by to nejrůznějším badatelům, anebo třeba i jiným studentům zaměřit se na různé otázky, jejichž analýza je pro ně jinak nemožná, protože sami nemají technické či finanční možnosti taková data nasbírat. Příkladem takové synergie je projekt Českého národního korpusu, který volným sdílením jazykových dat umožňuje lingvistům provádět empirický výzkum, aniž by museli věnovat úsilí sběru vlastního materiálu.

5. **Vzdělávání studentů**. Pro vzdělávání studentů je důležité, aby se s různými výzkumy seznamovali nejenom z publikovaných studií či knih, ale aby se rovněž naučili, jak je takový výzkum „udělán“. Sdílená data k tomu mohou významně dopomoci. Replikace výzkumu je například běžnou součástí výuky na zahraničních univerzitách – studenti se snaží zopakovat publikovaný výzkum na základě dat sdílených autory či na základě dat shromážděných vlastními silami při dodržování zásad daného výzkumu. Následně srovnávají výsledky, ke kterým dospěli, s původními závěry, hledají vysvětlení v případě, že se odlišují apod. Potřeba vypořádat se

s reálnými daty a konfrontace s kvalitním výzkumem pomáhá k odbornému růstu mnohem více než pouhé pasivní seznamování se s výsledky a hotovými interpretacemi.

6. **Společenská a morální zodpovědnost**. V České republice je naprostá většina výzkumu v oblasti jazyka a jeho užívání placena z veřejných peněz (institucionálních či grantových). Výsledky takto získané by tak měly být veřejně dostupné, a to nejen ve formě výsledných publikací (princip open access), ale i jako sdílená data a další materiály, které umožňují replikaci (viz iniciativa EOSC výše). Je dobré myslet na to, že data získaná ve výzkumu realizovaném v rámci takto placené práce nejsou z podstaty věci majetkem daného badatele či jeho hostitelské instituce, ale patří veřejnosti.

7. **Snadná dohledatelnost a zpětná interpretovatelnost**. Sdílení dat může být velmi užitečné i pro samotné badatele, kteří tato data dali dohromady. Nejeden ze čtenářů tohoto textu má jistě zkušenost se ztrátou vlastních dat, ať už to mělo jakékoliv důvody (například ztráta či znefunkčnění notebooku či úložného zařízení, výměna počítače, špatné zálohování apod.). Po několika letech od realizace určitého výzkumu může být navíc i pro samotného badatele obtížné datům rozumět. Pokud jsou data sdílena v rámci k tomu určené digitální platformy a jsou opatřena náležitou dokumentací (vč. metadat), má autor jistotu, že data kdykoliv dohledá a že bude schopen i po dlouhé době porozumět tomu, co datový soubor vlastně obsahuje. Využívání etablovaných a veřejně dostupných platforem (viz níže) navíc usnadňuje dohledatelnost a garantuje úroveň nezbytných metadat.

8. **Zveřejnění dat jako výstup**. V posledních letech se čím dál více prosazuje tendence nechápat jako výstupy vědeckého snažení jen určité publikace, ale také sdílená data. Datové soubory mohou být ošetřeny perzistentním digitálním identifikátorem (např. DOI nebo handle) a lze je citovat. Sdílení kvalitních dat by tak mělo být postaveno na roveň publikování kvalitních textů a badatelé, kteří sdílí data, by měli být za tuto aktivitu odměňováni stejně jako za aktivitu publikační.

9. **Metaanalýza**. Výše jsme zmínili to, že sběr dat bývá časově i finančně náročný proces. To má za následek, že v řadě případů není z praktického hlediska možné dosáhnout dostatečně rozsáhlého vzorku, aby si autoři mohli být jisti svými zobecněními (ve statistice se říká, že takové výzkumy mají nízkou statistickou sílu, a tedy hrozí, že zjištěné statisticky významné výsledky jsou ve skutečnosti nepřesné). Jsou-li dostupná data z různých výzkumů na stejné či podobné téma, mohou být analyzována souborně; takovým studiím se říká metaanalýzy. Zatímco jednotlivé dílčí studie jsou vystaveny různým zkreslením (plynoucím např. z použitého vzorku nebo stimulů apod.), souhrnná metaanalýza může pracovat se vzorkem participantů řádově větším, a může tak poukazovat na skutečné tendence, které z jednotlivých dílčích výzkumů nemusí být patrné. Příkladem velmi poctivé a náročné metaanalýzy může být studie Minny Lehtonenové et al. (2018), která se zaměřovala na dlouho diskutovanou hypotézu bilingvního zvýhodnění. Podle této myšlenky mají bilingvní

mluvčí oproti monolingvním mluvčím rozvinutější kognitivní kontrolu, tedy schopnosti, jako jsou pozornost či inhibice. Uvedená metaanalýza se snažila ověřit, zda je tato představa odůvodněná. Její autoři shromáždili 152 studií na toto téma (publikovaných i nepublikovaných) a v analýze tak pracovali s celkově 891 předchozími statistickými zjištěními. Na tomto základě přesvědčivě ukázali, že mezi bilingvismem a kognitivní kontrolou není žádný systematický vztah, tedy že nejspíše neplatí, že by bilingvní mluvčí byli oproti monolingvním nějak kognitivně zvýhodněni.

Příklady dobré praxe

Jak jsme již řekli na začátku, téma, které nastiňujeme v tomto článku, není pro české prostředí zcela nové. Jako příklady dobré praxe v našem prostředí můžeme zmínit Český národní korpus a dále infrastrukturu LINDAT/CLARIAH-CZ. Za účelem sdílení bylo vytvořeno několik platforem, níže stručně představíme dvě, s nimiž máme osobní zkušenost: Tromsø Repository of Language and Linguistics (viz TROLLing) a Open Science Framework (viz OSF).

Český národní korpus (viz ČNK), jako národní výzkumná infrastruktura se statusem K-centra evropské sítě Clarin, dnes slouží jako primární zdroj empirických dat o češtině. V praxi se zveřejňování dat, které v rámci projektu ČNK vzniknou, řídí politikou maximální otevřenosti – významným limitem při zveřejňování korpusu jsou však copyrightová omezení, kterými jsou zatíženy texty poskytované vydavateli či nakladateli. Korpusy ČNK jsou zveřejněny několika způsoby, především prostřednictvím specializovaných webových rozhraní (pro některé z nich je potřeba registrace a odsouhlasení licenčních podmínek zapovídajících komerční využití). Pro některé výzkumné účely to ovšem nestačí, ČNK je proto poskytuje v podobě speciálně upravených a licencovaných datových balíčků s promíchaným pořadím vět či odstavců, což znemožňuje zpětnou rekonstrukci originálu. Data, na něž se autorský zákon nevztahuje, jsou zveřejňována za výrazně volnějších podmínek, zejména prostřednictvím infrastruktury LINDAT/CLARIAH-CZ.

Důležitou součástí fungování ČNK je referenční platnost zveřejňovaných dat. Studie založené na korpusech, které jsou koncipovány jako referenční, tj. neměnné a veřejně dostupné (např. SYN2015 či ORAL2013), většinou vystačí s odkazem v bibliografii (bez nutnosti dalšího zveřejňování výzkumných dat), protože replikovatelnost je v rámci ČNK dostatečně zajištěna (leckdy i v rámci aplikací, které vytvářejí jednoznačné URL odkazující přímo k výsledkům).

Infrastruktura **LINDAT/CLARIAH-CZ** (viz LINDAT) je českým národním uzlem evropské sítě infrastruktur Clarin a Dariah. Vedle vývoje a provozování repozitáře nástrojů pro zpracování různých jazyků (taggery, parsery) poskytuje rovněž prostor pro sdílení datasetů (např. trénovacích dat) a korpusů. Primárně slouží repozitář pro větší projekty, které mají velký potenciál dalšího využití (nástroje a datasety vytvářené s ohledem na jejich širokou využitelnost). Vedle technicky kvalitně

zabezpečených repozitářů LINDAT jako clarinovské centrum typu B poskytuje také perzistentní identifikátory pro jednotlivé položky, kvalitní a federované vyhledávání v rámci sítě Clarin založené zejména na propracovaném a standardizovaném systému vytváření metadat.

Pro běžného badatele hledajícího platformu pro sdílení a zveřejňování výzkumných dat k jednotlivým studiím může být zajímavý projekt **Trolling** (viz TROLLing), který je – stejně jako předchozí – centrem sítě Clarin (jde o typ C, který se zaměřuje na sdílení metadat). Umístění výzkumných dat v rámci tohoto repozitáře umožní individuálním badatelům získat stabilní a perzistentní odkaz (DOI), který lze uvést do publikace. Díky zapojení do sítě Clarin a federovanému hledání jsou navíc takováto data přístupná širšímu spektru případných zájemců. Příkladem využití můžou být zdrojová data pro multidimenzionální analýzu češtiny (Cvrček et al., 2018) včetně skriptů v jazyce R pro jejich faktorovou analýzu (viz <https://doi.org/10.18710/QAJKZW>).

Oproti předchozím třem příkladům je **Open Science Framework** (viz OSF) platforma, která vzešla ze sociálněvědního prostředí a byla určitou reakcí na replikační krizi v psychologii. V současnosti je oborově velmi široce otevřená a nalezneme zde data a materiály z medicíny, přírodovědných, sociálních i humanitních oborů. Platformu spravuje Centrum pro otevřenou vědu (*Center for Open Science*), jehož posláním je zvyšovat otevřenost, integritu a reprodukovatelnost výzkumu. OSF nabízí uživatelsky příjemný a jednoduchý prostor pro sdílení celého procesu výzkumu. Nabízí možnost preregistrace experimentů a úložiště, v němž je možné sdílet data i nejrůznější další materiály. Jednotlivé projekty mohou zdarma získat identifikátor DOI, což je umožňuje snadno citovat v publikacích. Funkcí této platformy je větší množství. Patrně největší užitek z ní mohou mít badatelé, kteří se zaměřují na experimentálně zaměřený výzkum. Příkladem využití různých funkcí OSF může být projekt Jana Chromého *Comprehension of garden-path sentences in Czech* (<https://osf.io/dzcfe>), který obsahuje preregistrace, použité stimuly a výsledná data z experimentů.

Podněty pro české prostředí

Na závěr tohoto textu bychom chtěli vznést tři podněty pro české prostředí. Apelujeme v nich sice především na instituce, principiálně ale nic nebrání jednotlivcům se podle nich řídit už dnes. Jsme přesvědčeni, že jejich postupná implementace by výzkum jazyka a jeho užívání v českém prostředí ve střednědobém horizontu značně posunula, a zároveň si nejsme vědomi toho, že by někoho mohla poškodit, ba právě naopak: prospěch by z toho měli všichni zúčastnění (badatelé i studenti).

1. Akademické časopisy zaměřené na lingvistiku, které vychází v České republice, by měly přijmout dobrou praxi a od autorů studií, které uspěly v recenzním řízení, požadovat následné zveřejnění dat a dalších materiálů nutných k ověření

platnosti uvedených závěrů. Bez tohoto zveřejnění by nemělo být publikování textu možné.

2. Jednotlivá akademická pracoviště zaměřená na lingvistiku by měla po svých pracovnících požadovat sdílení dat a dalších výzkumných materiálů. Instituce by tak deklarovaly otevřenost, poctivost a transparentnost své vědecké činnosti.

3. Všechna akademická pracoviště, na nichž se realizují lingvisticky zaměřené studijní programy, by měla zavést jasná pravidla sdílení dat a dalších materiálů ze závěrečných prací (bakalářských, diplomových i disertačních) svých studentů.

Ve všech případech pochopitelně platí, že je třeba respektovat různá právní omezení (například nelze veřejně sdílet osobní údaje o účastnících výzkumů apod.). Při rozhodování o tom, která data zveřejnit, by kritériem měla být proveditelnost replikace, tj. zveřejnit to, co je pro její provedení nezbytné (týká se nejen dat, ale i metadat či metod). Běžně dostupné datasety či texty stačí uvést odkazem. Rovněž není nutné, aby byla data zveřejňována ihned po sesbírání. Dává smysl stanovit určitý časový odstup, který badateli či badatelskému týmu umožní realizovat bezprostředně plánované analýzy. Tento odstup by však neměl být delší než několik málo let.

Kriticky je třeba nicméně hodnotit praxi, kdy shromážděná data veřejně dostupná nejsou. Výzkum založený na neveřejných datech je v zásadě neverifikovatelný a jako takový ztrácí na důvěryhodnosti. Data, která by navíc mohla badatelské komunitě dál posloužit, a namísto toho jsou zamčena v soukromém archivu, zastarávají, ztrácejí na uplatnitelnosti, zvyšuje se riziko elektronické smrti formátu, v němž jsou uložena, a představují nejhorší možnou praxi nakládání s veřejnými zdroji i s intelektuálními kapacitami, které byly vloženy do jejich vybudování.

LITERATURA

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

CC ČR: *Creative Commons Česká republika* (2021). Cit. 16. 12. 2020. <https://www.creativecommons.cz>.

Cvrček, Václav – Komrsková, Zuzana – Lukeš, David – Poukarová, Petra – Řehořková, Anna – Zasina, Adrian Jan (2018): From extra- to intratextual characteristics: charting the space of variation in Czech through MDA. *Corpus Linguistics and Linguistic Theory* [online]. Cit. 15. 12. 2020. <10.1515/cllt-2018-0020>.

DHDC: Muñoz, Trevor – Flanders, Julia – Senseney, Megan (n. d.): *Digital Humanities Data Curation* [online]. Cit. 16. 12. 2020. <https://guide.dhcuration.org>.

EOSC: *European Open Science Cloud* (2021). Cit. 16. 12. 2020. <https://www.eosc-portal.eu>.

Chen, Jenn-Yeu (2007): Do Chinese and English speakers think about time differently? Failure of replicating Boroditsky (2001). *Cognition*, 104(2), s. 427–436.

January, David – Kako, Edward (2007): Re-evaluating evidence for linguistic relativity: reply to Boroditsky (2001). *Cognition*, 104(2), s. 417–426.

Lehtonen, Minna – Soveri, Anna – Laine, Aini – Järvenpää, Janica – de Bruin, Angela – Antfolk, Jan (2018): Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychological Bulletin*, 144(4), s. 394–425.

LINDAT: *LINDAT/CLARIAH-CZ: Czech Centre for Data Providing Certified Storage and Natural Language Processing Services* (2021). Cit. 16. 12. 2020. <https://lindat.cz>.

Nosek, Brian A. – Beck, Emorie D. – Campbell, Lorne – Flake, Jessica K. – Hardwicke, Tom E. – Mellor, David T. – van 't Veer, Anna E. – Vazire, Simine (2019): Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, 23(10), s. 815–818.

OSF: *Open Science Framework* (2011–2021). Center for Open Science. Cit. 16. 12. 2929. <osf.io>.

Shrout, Patrick E. – Rodgers, Joseph L. (2018): Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. *Annual Review of Psychology*, 69(1), s. 487–510.

TROLLing: *The Tromsø Repository of Language and Linguistics* (n. d.). Tromsø: The Arctic University of Norway. Cit. 16. 12. 2020. <https://dataverse.no/dataverse/trolling>.

Wiggins, Bradford J. – Christopherson, Cody Daniel (2019): The replication crisis in psychology: an overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, 39(4), s. 202–217.

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