LITERATURA

Elektronický slovník staré češtiny [online] (2006–). Praha: Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka. Cit. 15. 12. 2020. <https://vokabular.ujc.cas.cz>.

Staročeský slovník (1968–2008). Praha: Academia.

Vokabulář webový: Webové hnízdo pramenů k poznání historické češtiny [online] (2006–2021). Praha: Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka. Cit. 15. 12. 2020. <https://vokabular.ujc.cas.cz>.

Ústav pro jazyk český AV ČR, oddělení vývoje jazyka
Valentinská 91/1, 116 46 Praha 1
pytlikova@ujc.cas.cz, alenacerna@ujc.cas.cz, simek@ujc.cas.cz

Zpřístupňujme svá data, ale především chraňme konzultanty: stanovisko Pražské deskriptivní lingvistiky

Open your data, but protect your participants:
A statement by Prague Descriptive Linguistics

Veronika KAPIŠOVSKÁ, Jan KŘIVAN, Adam POSPÍŠIL, Florian SIEGL, Vít ULMAN, Jonáš VLASÁK, Veronika ZIKMUNDOVÁ, Jan ŽIDEK | Pražská deskriptivní lingvistika

Souhlasíme s tvrzením Václava Cvrčka a Jana Chromého, že empirický obrat v řadě odvětví lingvistiky přinesl zvýšený zájem o jazyková data, a to jak z hlediska jejich sběru a analýzy, tak z hlediska možností jejich publikování. Díky digitální revoluci je vědecká práce s daty výrazně snazší než dříve a existují také nové způsoby, jak data archivovat i zpřístupňovat širšímu publiku. Digitalizace však nemá sama o sobě přímý vliv na vytváření primárních jazykových dat a na otázky s tím spojené, zvláště pokud jde o data z terénního výzkumu. Vzhledem k tomu, že tento rozměr bývá přehlížen a hrozí, že by i v této diskusi mohl zůstat opomenut, představíme ve svém příspěvku několik zkušeností z dřívější doby analogové, které mohou zveřejňování dat do jisté míry zkomplikovat.

Typ dat, jimiž mají v plánu přispívat členové naší skupiny[1] do online repozitářů jazykových materiálů v režimu open access, jsou jazykové korpusy různých, dosud málo zdokumentovaných a ohrožených jazyků. Takové korpusy však nelze jednoduše srovnávat s korpusy velkých jazyků. Příkladem může být tvorba korpusu mluveného jazyka. I v případě standardních evropských jazyků, jako je čeština, se jedná o časově náročný proces, protože musí být mluvená podoba jazyka za pomoci jasně daných pravidel převedena do psaného formátu. Zvláštním případem je však v tomto

[1] Pražská deskriptivní lingvistika (<http://linguistics.cz/about-us>).

kontextu korpus málo dokumentovaného, ohroženého, převážně mluveného jazyka vytvářený terénním lingvistou ve spolupráci s konzultanty.[2]

Pro ilustraci uvedeme zkušenost jednoho z členů naší skupiny. Vytvoření hrubého transkriptu jedné minuty mluveného jazyka (monologu/vyprávění) se čtyřmi rovinami anotace a překladem trvá v programu ELAN přibližně tři hodiny (z toho asi jednu hodinu trvá transkripce a překládání společně s konzultantem, další dvě hodiny pak manuální anotace v programu ELAN). To znamená, že anotace deseti minut projevu, která vede k první verzi transkriptu, může zabrat asi týden práce.[3] Transkripty je pak potřeba před zveřejněním cíleně ověřovat u rodilého mluvčího, v případě ohrožených jazyků většinou během dalších výjezdů výzkumníka do terénu.[4] Objem zpracovaných dat pak odpovídá tomu, že většinu této práce dělá obvykle jen jeden výzkumník (v lepším případě menší skupina). Malý počet lingvistů dokumentujících ohrožené jazyky je dán nejen obtížností práce s nimi, ale také nízkým odborným zájmem o výstupy; rozsahem malý soubor anotací neumožňující kvantitativní analýzu také dnes většinou není považován za hodnotný vědecký výstup, ačkoliv už tato fáze výzkumu je založena na analýze dat ukotvené v adekvátním teoretickém rámci.

Výše uvedené otázky spojené s přípravou dat pro publikování jsou však jen jednou stranou mince. Ne všechna data totiž mohou být zpřístupněna, např. z důvodu zachování anonymity. Potřeba jejího zajištění se vztahuje k několika rovinám: už jen samotný osobní vztah mezi výzkumníkem a konzultantem určuje, o čem se může mluvit a co může být nahráváno. Zatímco části etnograficky, historicky nebo osobně citlivých projevů mohou být využity ve výzkumu (např. větné příklady), projevy s citlivým obsahem jako celek již nikoli. Přístup k takovým datům proto musí být omezen (tento problém mohou vyřešit různé technické nástroje v online repozitářích). V krajním případě může být nutné anonymizovat jména konzultantů a/nebo místo nahrávání, aby byli konzultanti a jejich rodiny chráněni před identifikací a případnou politickou perzekucí, a to i tam, kde to samotný obsah vyprávění nevyžaduje. V tomto ohledu má výzkumník dvojí zodpovědnost: jako vědec je zodpovědný za zveřejnění výsledků výzkumu, stejně tak je ale zodpovědný za bezpečí svých konzultantů a jejich rodin, přičemž tato zodpovědnost začíná už na samém počátku dokumentace.

Lze předpokládat, že internetové připojení bude v průběhu 21. století dostupné i v nejodlehlejších oblastech světa. Data týkající se ohrožených jazyků (jimiž se často hovoří daleko na periferii) publikovaná podle principu „FAIR“ (Cvrček – Chromý, s. 6) by proto měla být vhodným způsobem přístupná i daným jazykovým komunitám. Jejich mluvčí však nejsou odborníky a někdy nehovoří ani majoritním jazykem své země, natož jazykem světovým. Je proto potřeba zajistit, aby na to rozhraní online repozitářů byla připravena.

---

[2] Na tomto místě je třeba zmínit, že značné množství (nejen) ohrožených jazyků postrádá lingvisticky spolehlivé standardy pro přepis do psané podoby. Např. oficiální pravopis hauštiny (jistě ne ohroženého jazyka) neznačí ani tón, ani vokalickou délku, ačkoliv jsou oba tyto jevy fonologicky relevantní.

[3] Tuto práci však nelze provádět v prvních fázích terénního výzkumu, neboť jejím předpokladem je hluboká strukturní znalost jazyka a také dobrá spolupráce s rodilým konzultantem. Snížením počtu anotačních rovin (fonetická transkripce, fonologická reprezentace, slovnědruhová analýza, morfologické značkování) se může rychlost zpracování samozřejmě částečně zvýšit.

[4] Běžně se stává, že badatel počáteční transkript v průběhu let zásadně přehodnotí, což ztěžuje jeho rychlé publikování. V této souvislosti bychom chtěli zdůraznit, že výše popsaný transkript, který lze v programu ELAN otevřít společně s časově zarovnaným zvukovým souborem, zcela naplňuje standardy transparentních dat; je naším přáním, aby se uživatelé nespokojili jen se sekundárním transkriptem, ale aby využívali možnosti poslechnout si zdrojovou nahrávku a samostatně hodnotit navrženou analýzu.

Tento text by neměl být interpretován jako odmítnutí pozic úvodního příspěvku. Plně se hlásíme ke stanovisku, že data musí být zpřístupněna, aby umožnila ověřování a replikaci výzkumných výsledků, nicméně – jak jsme se pokusili nastínit –, parametry publikace na jedné straně závisí na zpracování primárních dat, na druhé straně nejsou jen otázkou lingvistiky samotné. Máme zodpovědnost jak vůči komunitě vědecké, tak vůči komunitě mluvčích daného jazyka.

Pražská deskriptivní lingvistika
prague.descriptive@linguistics.cz
www.linguistics.cz

K otevřenosti dat využívaných v onomastice

On open data in onomastics

Pavel ŠTĚPÁN, Soňa WOJNAROVÁ | Ústav pro jazyk český AV ČR, oddělení onomastiky

Onomastika patří mezi ty lingvistické disciplíny, v nichž empirická data sehrávají zcela zásadní roli. Je však třeba zdůraznit, že vzhledem ke značné pestrosti onomastických výzkumných témat, orientovaných jak synchronně, tak diachronně, jsou využívána data nejrůznějšího charakteru. To platí jistě nejen o onomastice, ale i o dalších oblastech jazykovědy. Jeví se proto jako poněkud obtížné a ne zcela adekvátní nastavovat jednotná měřítka, nebo dokonce závazná pravidla pro zveřejňování dat nejrůznějšího charakteru a původu.

Dlouhodobým úkolem oddělení onomastiky Ústavu pro jazyk český AV ČR je zpracování Slovníku pomístních jmen v Čechách (2014–2021; Matúšová et al., 2005–2009). Přípravné práce na tomto díle započaly již v 60. letech 20. století. V letech 1963–1980 proběhla rozsáhlá soupisová akce, během níž bylo díky účasti několika set dobrovolných místních spolupracovníků nashromážděno více než 400 000 pomístních jmen z území Čech (více viz Olivová-Nezbedová – Malenínská, 2000). Tato sbírka pomístních jmen není dostupná online (a z finančních a personálních důvodů se s jejím zpřístupněním touto formou zatím nepočítá), avšak po dohodě je k dispozici k nahlédnutí všem badatelům, kteří mohou její data využívat pro jakékoli vlastní výzkumy. Možné je samozřejmě i použití těchto dat pro případnou replikaci výzkumů již publikovaných. Vzhledem k tomu, že v současné době vznikají paralelně dvě samostatná slovníková díla − vedle zmíněného Slovníku pomístních jmen v Čechách je v dialektologickém oddělení Ústavu pro jazyk český AV ČR v Brně zpracováván podle částečně odlišné koncepce Slovník pomístních jmen na Moravě a ve Slezsku (2014–2021) − je velmi důležité, aby obě pracoviště v plném rozsahu zpřístupňovala svá data (ať už v jakékoli podobě) ostatním badatelům zabývajícím se výzkumem pomístních jmen, zvláště pak kolegům pracujícím na paralelním projektu (důležitost srovnávacích dat z jiného území je nesporná i kvůli vytvoření celkového obrazu pomístních jmen na území Česka).

Mezi další úkoly oddělení onomastiky patří ověřování podob jmen a příjmení z hlediska zákonných ustanovení usměrňujících jejich zápis do osobních dokladů. Od roku 2009 bylo (do 31. 12. 2020) vystaveno celkem 1 424 posudků rodných jmen i příjmení. Vystavené posudky a žádosti