KORPUS POLSKÉHO ZNAKOVÉHO JAZYKA

Ve dnech 28. 2. – 2. 3. 2018 proběhlo setkání pracovníků Ústavu jazyků a komunikace neslyšících a Ústavu obecné lingvistiky Filozofické fakulty Univerzity Karlovy s týmem připravujícím korpus polského znakového jazyka (polski język migowy: PJM) v rámci Kabinetu lingvistiky znakových jazyků na Varšavské univerzitě. Náplní setkání bylo jednak seznámení s aktuálním stavem příprav korpusu PJM, jednak jednání o možné budoucí spolupráci mezi pražskými a varšavskými lingvisty znakových jazyků, a to především s perspektivou možného vzniku korpusu českého znakového jazyka (ČZJ).

Polský korpusový projekt představuje bezesporu pozoruhodný počin, nejen v kontextu střední a východní Evropy, tedy oblasti, kde je lingvistika znakových jazyků prozatím relativně málo rozvinutým oborem, ale i z celosvětového hlediska. Týmu pod vedením Pawła Rutkowského se podařilo během osmi let sestavit jeden z největších korpusů znakového jazyka,[1] a to v podstatě „na zelené louce", tj. aniž předtím existovala zásadnější tradice popisu a lingvistického poznání PJM.[2] V současné době korpus obsahuje vzorky produkce PJM pocházející od více než stovky mluvčích v celkové délce přesahující 400 hodin a zároveň komplexní anotaci velkého množství jazykových rovin. Zveřejnění korpusu je plánováno na rok 2019.

Ve svých počátcích vycházel projekt metodologicky z korpusu německého znakového jazyka (Deutsche Gebärdensprache: DGS, Prillwitz et al., 2008), který začal vznikat na Hamburské univerzitě v roce 2009 a coby jeden z prvních znakových korpusů posloužil jako inspirace i pro další projekty. Ačkoli v současnosti probíhá (či se rozbíhá) práce na korpusech vícera znakových jazyků, dosud existuje pouze hrstka dokončených a (alespoň částečně) veřejně dostupných zdrojů tohoto typu. Kromě korpusu DGS jsou to především korpusy nizozemského znakového jazyka (Nederlandse Gebarentaal: NGT, Crasborn, Zwitserlood, & Ros, 2008), britského znakového jazyka (British Sign Language: BSL, Schembri, Fenlon, Rentelis, & Cormier, 2017) a australského znakového jazyka (Australian Sign Language: Auslan, Johnston, 2010).

Vedle P. Rutkowského se pražská skupina setkala s Joannou Filipczak, Annou Kuder a Piotrem Mostowským, kteří tvoří jádro varšavského korpusového týmu. J. Filipczak prezentovala aktuální postup práce na anotaci nahrávek, což byl — spolu s představením metodologie sběru materiálu — hlavní diskusní bod celého setkání. Níže shrnujeme stěžejní informace k oběma bodům.

SBĚR MATERIÁLU

Aby mohl být korpus PJM využíván jako referenční zdroj pro široké spektrum lingvistických studií, musí materiál v něm obsažený zachycovat diverzifikované jazykové projevy produkované co možno nejreprezentativnějším vzorkem neslyšících mluv-

1 A nutno dodat, že také jeden z největších multimodálních korpusů obecně, tedy včetně mluvených korpusů, které obsahují videonahrávky.

2 Viz např. Rutkowski & Sak, 2016.

čích, jejichž prvním jazykem je PJM. Pro zachycení regionální variace byl počet mluvčích stanoven kvótně podle proporce populace v jednotlivých vojvodstvích. Mluvčí jsou dále rovnoměrně rozděleni do věkových skupin (v rozpětí 18 až 92 let), přičemž jsou (resp. ve finálním vzorku budou) stejně zastoupena obě pohlaví. Další metadata o informantech[3] byla získávána prostřednictvím dotazníku.

Samotné nahrávání probíhalo jednak v nahrávacím studiu ve Varšavě, jednak přímo v regionech ve speciálně upraveném autobuse. V úplné konfiguraci tvořila nahrávací zařízení pětice HD kamer. Jednotlivých sezení se účastnili vždy 2 informanti a 1 neslyšící moderátor. Procedura sběru materiálu sestávala ze série úloh zaměřených na elicitaci různých typů jazykové produkce. Elicitační úlohy lze rozdělit do několika skupin: (1) individuálně produkované narativy — převyprávění grafických, filmových a znakovaných příběhů, sloužící k získání srovnatelného materiálu od všech informantů, (2) interakční produkce s různou úrovní spontaneity (od řízené diskuse po volnou konverzaci), tj. produkci s vyšší mírou ekologické validity, umožňující např. studium konverzačních fenoménů, (3) individuální produkce izolovaných znaků pro lexikografické účely. Celkem se jednalo až o 24 úloh, reálná čísla ale u individuálních mluvčích kolísají.[4] V průběhu let pořizování materiálu se některé úlohy vyřazovaly, protože se ukázalo, že nenaplňují původní záměr, jsou např. příliš obtížně formulované, informanti na ně reagují omezeně nebo stereotypně. Některé úlohy se naopak připojovaly, a to za účelem získání konkrétního typu dat, např. popisu prostoru.

Jednotlivá nahrávání trvala s přestávkami až 5 hodin, přičemž informanti obdrželi za svou účast malou finanční kompenzaci.

ANOTACE

Polský korpus se řadí mezi největší projekty svého druhu nejen rozsahem nasbíraného materiálu, ale především komplexně pojatou anotací. Korpus PJM se odlišuje od jiných korpusů (NGT, BSL, AUSLAN a množství dalších, menších projektů) především volbou softwarového nástroje pro anotaci videonahrávek. Přestože běžně využívaným programem pro anotaci multimodálního materiálu je *ELAN* (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006), polský tým zvolil program *iLex* (Hanke, Storz, & Wagner, 2010). Tento program byl vyvinut přímo pro anotaci korpusových dat v rámci korpusu DGS, je tudíž vybaven většinou potřebných funkcí pro zpracovávání tohoto specifického materiálu. Jeho hlavní výhodou je však možnost skupi-

3 V případě korpusů znakových jazyků je zásadní informací mimo jiné jazykové pozadí mluvčího. Status „rodilý mluvčí" může být problematický mimo jiné proto, že kolem 90 % neslyšících vyrůstá ve slyšících rodinách — označení rodilý mluvčí je tedy nutno chápat šířeji než u mluvčích mluvených jazyků (srov. např. Costello, Fernández, & Landa, 2008). Další důležitou informací je charakter a lokalita vzdělávací instituce, protože míra vystavení znakovému jazyku se výrazně liší v závislosti na době školní docházky i podle jednotlivých škol.

4 Od každého mluvčího se podařilo získat data ze základního souboru 11 úloh.

nové spolupráce velkého množství anotátorů pracujících s materiálem deponovaným v cloudovém úložišti a možnost zpětných plošných úprav v celém objemu anotovaného materiálu. To je klíčová výhoda[5] vzhledem k tomu, že na anotaci polského souboru korpusu kontinuálně pracuje 11 neslyšících anotátorů (pod vedením neslyšícího supervizora), poněvadž tím odpadá nutnost logisticky náročné distribuce velmi objemných mediálních souborů mezi jednotlivé anotátory.

Základní úrovní anotace jsou glosy zastupující formu lexikálních jednotek, tj. tagy reprezentující kombinace fonologických rysů (tvar, orientace a umístění ruky, směr pohybu — nikoli však nemanuální komponenty znaků), doplněné notací podle systému HamNoSys (Hanke, 2004) a polským překladem. Základní fonologická informace je na rovině glos rozšířena o další parametry: obouruční a simultánní artikulaci, charakter deixe, reduplikaci, kompozici a další. Na lexikální úrovni[6] obsahuje korpus téměř 500 000 tokenů a přes 6000 lemmat. Další anotované roviny zahrnují slovní druhy (s více než 30 kategoriemi), pohyby hlavy a těla, manuální negaci a také základní syntaktickou anotaci na úrovni tzv. „clause-like units" (srov. Hodge, 2013).

Jak je patrné, podoba anotace nad rámec glos vychází ze specifických výzkumných otázek, podobně jako je tomu u jiných korpusů znakových jazyků. Co nejúplnější anotace na rovině glos je pro využitelnost každého korpusu znakového jazyka jako základního nástroje pro základní deskriptivní studie (tj. popis frekvenčních a distribučních charakteristik lexika) zcela zásadní. Nadto je anotace glos nutnou podmínkou pro anotaci dalších rovin.

VÝHLED: KORPUS ČESKÉHO ZNAKOVÉHO JAZYKA

Ačkoli polský korpus dosud nebyl zveřejněn, vzniklo na jeho základě již několik deskriptivních a analytických studií o PJM (Rutkowski, Kuder, Czajkowska-Kisil, & Łacheta, 2015; Rutkowski & Łozińska, 2016). Lze předpokládat, že impakt, který bude mít vznik referenčního korpusu na polskou lingvistiku znakových jazyků, bude značný. Je lehce paradoxní, že ačkoli polská lingvistika znakových jazyků nemá tak dlouhou tradici a institucionální zakotvení jako ta česká, nachází se nyní díky korpusu PJM v mnohem lepší situaci jak co do možností vlastní vědecké práce, tak co do mezinárodního ohlasu a spolupráce.[7]

Vznik korpusu českého znakového jazyka představuje pro českou lingvistiku znakových jazyků aktuálně možná největší desideratum. Cesta k budoucímu korpusu, byť bude bezesporu dlouhá, je však otevřená — přinejmenším co se týče technického

5 Zároveň však iLex postrádá mnoho výhod ELANu, především horizontálně synchronizované rozhraní, otevřený kód a s ním spojenou možnost integrace dalších programů a aplikací. Další nevýhodou iLexu je to, že neumožňuje snadné vyhledávání na základě regulárních výrazů, naopak vyžaduje poměrně pokročilé programování.

6 Tj. lexikální znaky, klasifikátorové konstrukce a jmenné znaky. Kromě toho jsou zvlášť anotována gesta a tzv. kulturní znaky.

7 Kromě zmíněného Hamburku spolupracuje varšavský tým intenzivně také s Trevorem Johnstonem, předním odborníkem na auslan.

vybavení a infrastruktury, jimiž disponuje pražský Ústav jazyků a komunikace neslyšících. Úzká spolupráce s varšavskými kolegy může být na této cestě zásadním benefitem.

LITERATURA:

Costello, B., Fernández, J., & Landa, A. (2008). The non- (existent) native signer: sign language research in a small deaf population. In R. M. de Quadros (Ed.), *Sign Languages: spinning and unraveling the past, present and future. TISLR9* (pp. 77–94). Petrópolis: Editora Arara Azul.

Crasborn, O., Zwitserlood, I., & Ros, J. (2008). *The Corpus NGT. A digital open access corpus of movies and annotations of Sign Language of the Netherlands*. Nijmegen: Centre for Language Studies, Radboud Universiteit Nijmegen.

Hanke, T. (2004). HamNoSys — Representing Sign Language Data in Language Resources and Language Processing Contexts. In O. Streiter & C. Vettori (Eds.), *LREC 2004, Workshop proceedings: Representation and processing of sign languages* (pp. 1–6). Paris: ELRA.

Hanke, T., Storz, J., & Wagner, S. (2010). iLex: Handling Multi-Camera Recordings. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies* (pp. 110–111). Paris: ELRA.

Hodge, G. (2013). *Patterns from a signed language corpus: Clause-like units in Auslan (Australian sign language)* (nepublikovaná dizertační práce). Macquarie University, Sydney.

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, *15*(1), 106–131.

Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., & Schwarz, A. (2008). DGS Corpus Project-Development of a Corpus Based Electronic Dictionary German Sign Language / German. In O. Crasborn, E. Efthimiou, E. D. Thoutenhoofd, & I. Zwitserlood (Eds.), *LREC 2008 Workshop Proceedings. W 25: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora* (pp. 159–164). Paris: ELRA.

Rutkowski, P., Kuder, A., Czajkowska-Kisil, M., & Łacheta, J. (2015). The structure of nominal constructions in Polish Sign Language (PJM): A corpus-based study. *Studies in Polish Linguistics*, *10*(1), 1–15.

Rutkowski, P., & Łozińska, S. (2016). Argument Linearization in a Three-Dimensional Grammar: A Typological Perspective on Word Order in Polish Sign Language (PJM). *Journal of Universal Language*, *17*(1), 109–134.

Rutkowski, P., & Sak, M. (2016). Sign Language: Eastern Europe. In G. Gertz & P. Boudreault (Eds.), *The SAGE Deaf Studies Encyclopedia* (pp. 796–798). Thousand Oaks, CA: SAGE Publications, Inc.

Schembri, A., Fenlon, J., Rentelis, R., & Cormier, K. (2017). *British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2017 (Third Edition)*. London: University College London.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556–1559). Paris: ELRA.

Jakub Jehlička
Hana Prokšová
Lucie Stanovská