ČASOPIS PRO MODERNÍ FILOLOGII 106, 2024, Č. 1, S. 107–109

PRVNÍ KORPUS MLUVČÍCH ČEŠTINY V DĚTSKÉM VĚKU

Korpusy mluveného i psaného jazyka hrají klíčovou roli ve všech lingvistických disciplínách. Z dosavadní české lingvistické praxe jsou dobře známé rozsáhlé obecné korpusy i různé druhy specializovaných databází, sdružené především pod hlavičkou Českého národního korpusu [korpus.cz]. Český národní korpus spravuje také přístup do akvizičních korpusů vzniklých pod značkou AKCES [akces.ff.cuni.cz], zaměřených na produkci dětí ve školním vyučování, které si osvojují češtinu jako rodilí i nerodilí mluvčí. Korpus Chroma, publikovaný v červenci 2023 péčí badatelského týmu CoCzeFLA [coczefla.ff.cuni.cz] v rámci velké výzkumné infrastruktury LINDAT [lindat.cz, https://lindat.cz/], jde chronologicky ještě kousek dál a zachycuje produkci monolingvních dětí osvojujících si češtinu v dětském věku (mezi 19 až 49 měsíci).

Celosvětový výzkum akvizice prvního jazyka v dětském věku má pro svá korpusová bádání k dispozici unikátní nástroj, mezinárodní databázi CHILDES [childes.talkbank.org], v níž je zařazen i korpus Chroma. Tento korpus je tak třeba vnímat jednak v kontextu dostupných českých akvizičních korpusů dětí školního věku, kde zaplňuje mezeru chronologickou, jednak v kontextu jinojazyčných korpusů v databázi CHILDES, kde zaplňuje mezeru typologickou a hraje významnou diverzifikační roli. CHILDES totiž aktuálně zahrnuje 75 korpusů zachycujících monolingvní osvojování angličtiny, což tvoří asi 28 % této databáze, zatímco celá skupina slovanských jazyků je reprezentovaná pouhými 11 příspěvky včetně korpusu Chroma (4 %).

Korpus Chroma je založen na audionahrávkách spontánních mluvených interakcí dětí s blízkými pečovateli. Nahrávky všech sedmi zapojených dětí byly pořizovány longitudinálně v pravidelných intervalech a korpus tak pro každé z nich zachycuje 11 až 27 měsíců vývoje s průměrným objemem nahrávek 36 minut za měsíc. Při nahrávání nebyl přítomen žádný badatel. Každá rodina dostala k dispozici nahrávací zařízení a několik základních pokynů: nenahrávat více než dvě dospělé osoby najednou, omezit hluk v pozadí a zachycovat běžné každodenní situace. Typickými zachycenými situacemi jsou hraní s hračkami, prohlížení knížek, stolování a domácí práce. Nahrávky samotné nejsou veřejné, je ale možné požádat si přes web CoCzeFLA o zpřístupnění některých z nich pro další analýzu. Veřejná část korpusu sestává z přepisů těchto nahrávek. Jde celkem o 183 textových souborů; každá nahrávka, resp. všechny nahrávky pořízené během jednoho dne od jednoho dítěte, má vlastní soubor přepisu označený identifikátorem dítěte, jeho přesným věkem v době pořízení nahrávky a několika dalšími údaji. Korpus celkem obsahuje 99 388 tokenů v 42 103 dětských promluvách a 238 211 tokenů v 61 252 promluvách dospělých.

Přepisy korpusu Chroma se řídí systémem přepisu CHAT [talkbank.org/manuals/CHAT.pdf], což je odlišuje od jiných dostupných českých akvizičních korpusů. Systém CHAT sjednocuje všechny korpusy zahrnuté v databázi CHILDES. Přepisy jsou čitelné v jakémkoliv textovém editoru. Klíčovými prvky jsou (1) členění na řádky po jednotlivých výpovědích, (2) rozlišování hlavních a vedlejších řádků a (3) označování mluvčích.

ISSN 2336-6591

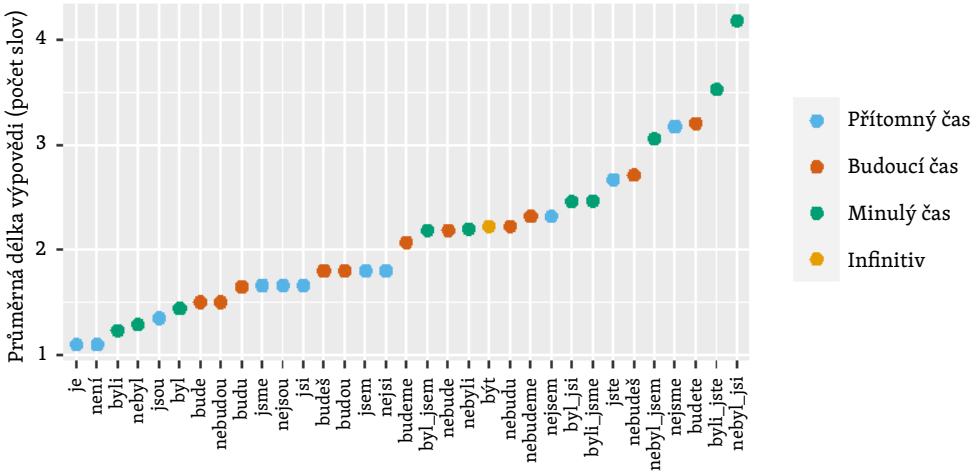
Ukázka: *CHI: tam byla hasiče [*] .
%xmor: adv:pro|tam v|být-x&past&akt&x&impf-err n|hasič-x&x&MA-err .
%xpho: tam byla hasiše .
%err: byla = byli , hasiče = hasiči .

Hlavní řádek začíná znakem * a identifikací mluvčího (CHI ‚child' vždy značí centrální nahrávané dítě a k němu se vztahují ostatní účastníci: MOT ‚mother', BRO ‚brother' apod.). Tento řádek obsahuje ortografický přepis, v němž se častokrát odhlíží od artikulačních nedokonalostí dětí v raných fázích osvojování, pokud se ale dětská realizace odchyluje významně, zachycuje tuto odchylku i hlavní řádek (např. *manyky* jako nedokonalý pokus o cílovou formu *mandarinky*). Takové případy jsou označené speciálním kódem (*manyky@c*). Svůj kód mají např. i tvarotvorné a slovotvorné inovace (*sáňkovají@n* místo *sáňkují*; *čokoládovář@n* jako ten, kdo vyrábí čokoládu) a další jevy typické pro dětskou a na děti orientovanou produkci.

K hlavnímu řádku se vztahují následující vedlejší řádky uvozené znakem %. V ukázce výše jde o řádky xpho, err a xmor. Řádek xpho obsahuje velmi přibližný fonologický přepis. Vedle českých ortografických znaků využívá navíc znak @ pro šva. Řádek err obsahuje specifikaci chyby signalizované kódem [*] na hlavním řádku. Řádek xmor obsahuje morfologickou anotaci. Ta byla provedena pomocí volně dostupného českého nástroje MorphoDiTa [ufal.mff.cuni.cz/morphodita]. Úspěšnost samotné automatické anotace byla na velmi vysoké úrovni, navíc byly ale provedeny i ruční zásahy a opravy (viz detailní popis v článku Chromá et al., 2024). Kromě toho byla anotace MorphoDiTou na závěr převedena do formátu zohledňujícího podobu morfologického tagu v databázi CHILDES (viz ukázka výše).

Díky morfologické anotaci je velmi snadné korpus prohledávat nejen podle jednotlivých tvarů, ale rovněž podle lemmat, slovních druhů, morfologických kategorií nebo jejich kombinací ve výpovědi. V rámci platformy CHILDES je k dispozici specializovaný software CLAN určený jak přímo pro přepis, tak i pro analýzu přepisů ve formátu CHAT. Tento software umí nejen vyhledávat podle zadaných dotazů a spočítat frekvenci či průměrnou délku výpovědi (což je klíčové měřítko ve výzkumu jazykového vývoje), ale je určen i pro náročnější automatizované úlohy. Umí například využít morfologickou anotaci k posouzení úrovně produktivity dítěte při užívání vybraných morfosyntaktických jevů prostřednictvím měřítka Index produktivní syntaxe (viz Matiasovitsová et al. 2023). Výhodou CLANu je, že předem počítá se strukturou CHATovského přepisu, např. s označováním typů řádků, kódováním nesrozumitelných částí nebo i s formátem morfologické anotace. To ale může být chápáno i jako nevýhoda, protože určitá část vyhledávání se děje automaticky a může zůstat nepovšimnutá (odfiltrování výpovědí s nesrozumitelnými částmi) a také je v dotazech nutné zohledňovat specifickou podobu české anotace, která odráží morfologické vlastnosti češtiny. Kromě CLANu je ale možné použít i jiný software (např. NotePad++) umožňující hledání ve větším množství textových souborů najednou na základě regulárních výrazů.

V rámci týmu CoCzeFLA už probíhají nebo se připravují analýzy zaměřené například na vývoj reference k mluvčímu a adresátovi pomocí první a druhé slovesné a zájmenné osoby i pomocí třetí osoby (*maminka pofouká*); na morfologicky nebo syntakticky inovativní dětskou produkci (*psu@n* = píšu, *kolotočí@n* = točí se na kolo-



GRAF 1. První výskyt tvarů slovesa být ve vztahu k průměrné délce výpovědi

toči); nebo na vývoj tvarové diverzity sloves a jmen. Jedním z rozpracovaných témat je analýza posloupnosti tvarů slovesa *být* během osvojování, jak ji zachycuje korpus CHROMA (viz Graf 1 výše). Vedle zkoumání různých morfosyntaktických jevů se však korpus hodí také pro analýzy slovotvorné, lexikální nebo třeba konverzační.

Výzkum osvojování prvního jazyka by měl v optimálním případě být založen na komplementárním využívání experimentálních a korpusových dat. Vydání korpusu CHROMA jakožto prvního takto zaměřeného českého korpusu je proto klíčovým počinem pro další výzkum osvojování češtiny.

LITERATURA

CHROMÁ, A. et al. (2024): A morphologically annotated longitudinal corpus of spoken Czech child-adult interactions. *Language Resources and Evaluation*, s. 1–24.

MATIASOVITSOVÁ, K. et al. (2023): The Validity of a Transcript-Based Measure of Child Language Development in Czech. In: P. Gappmayr, — J. Kellogg (eds.), *BUCLD 47: Proceedings of the 47th Annual Boston University Conference on Language Development. Boston: Cascadilla Press*, s. 533–547.

Anna Chromá | Ústav obecné lingvistiky & Ústav českého jazyka a teorie komunikace, Filozofická fakulta Univerzity Karlovy | nám. Jana Palacha 2, 116 38 Praha 1
ORCID ID: 0000-0003-0559-076X
anna.chroma@ff.cuni.cz

Klára Matiasovitsová | Ústav českého jazyka a teorie komunikace, Filozofická fakulta Univerzity Karlovy | nám. Jana Palacha 2, 116 38 Praha 1
ORCID ID: 0000-0002-2338-070X
klara.matiasovitsova@ff.cuni.cz