Rodinám zapůjčíme kameru nebo diktafon a nahráváme spontánní dětskou komunikaci, popisuje tvorbu Korpusu rané češtiny Anna Chromá

O významu Korpusu vývoje řeči českých dětí hovoří v rozhovoru iniciátorka projektu, kurátorka dat a doktorandka obecné lingvistiky Anna Chromá. Jedná se o jediný korpus svého druhu pro češtinu a jako takový má zásadní význam pro popis průběhu osvojování češtiny u malých dětí. Další využití má v pedagogice, logopedii, ale i pro vývoj diagnostických a terapeutických nástrojů pro děti s jazykovými poruchami.

Korpus je součástí projektu LINDAT/CLARIAH-CZ řešeným na FF UK a také nově vzniklého Centra pro digital humanities FF UK.

Korpus rané češtiny je jediný svého druhu pro český jazyk. Na co přesně se v něm zaměřujete?

Momentálně se soustředíme na dlouhodobé nahrávání několika málo dětí, které začínáme sledovat ideálně v době, kdy produkují pouze jednoslovné výpovědi, kolem 18. měsíce věku dítěte. Nahráváme jejich spontánní komunikaci s rodičem, vždy přibližně hodinu každé tři týdny, a nahrávky přepisujeme, takže pro každé dítě vznikne balík anonymizovaných přepisů zachycujících jeden až dva roky jeho vývoje.

Jakým způsobem korpus budujete?

Zapojeným rodinám zapůjčíme kameru nebo diktafon a probereme s nimi základní požadavky na nahrávání. Mezi ně patří pravidlo nenahrávat příliš mnoho lidí najednou a omezit hluk v pozadí. Jinak by se ale rodiče měli snažit zachycovat spontánní situace, jak by se odehrály i bez nahrávacího zařízení, a dodržovat intervaly mezi nahrávkami. Pak už je to na nich.

Pokud vzniká větší pauza mezi nahrávkami, tak se rodině

připomeneme a snažíme se ji motivovat. Ale celá účast je založena výhradně na dobrovolné spolupráci. Nahrávání pak probíhá ideálně po dobu dvou let. Některé rodiny ale skončí třeba po roce nebo roce a půl, třeba když se narodí mladší sourozenec, a organizace nahrávání se stává obtížnější.

Je náročné pro tento projekt najít vhodné rodiny? Jaké podmínky musí taková rodina splnit a hledáte aktuálně i nové „přispěvatele“?

Nahrávající rodiny jsme dosud vždycky hledali mezi přáteli a kolegy. Součástí korpusu je i má vlastní rodina a víc než polovina ze čtrnácti dlouhodobě sledovaných rodin má nějaký vztah k FF UK, ať už jako studující, zaměstnanci nebo absolventi. Náš vzorek je malý a nemůžeme si dovolit na něm zkoumat faktory jako vzdělání nebo socioekonomický status rodičů, takže z tohoto hlediska ho budujeme spíš homogenně.

Zároveň zatím zaznamenáváme pouze vývoj dětí z jednojazyčného českého prostředí, které se vyvíjí typicky, tedy bez jazykových, smyslových nebo intelektových vývojových poruch. Takže vhodných rodin, které splňují výzkumné požadavky, je dostatek. Obtížnější je samozřejmě splnit nároky na vytrvalost v nahrávání a spolehlivost v intervalech. Tam je klíčová vnitřní motivace našich účastníků. Taky jde o zásah do soukromí, zejména u videonahrávek, a i když nahrávku v zásadě vidí jen dvě osoby – přepisující a revidující – naši účastníci s tím musí být srozuměni.

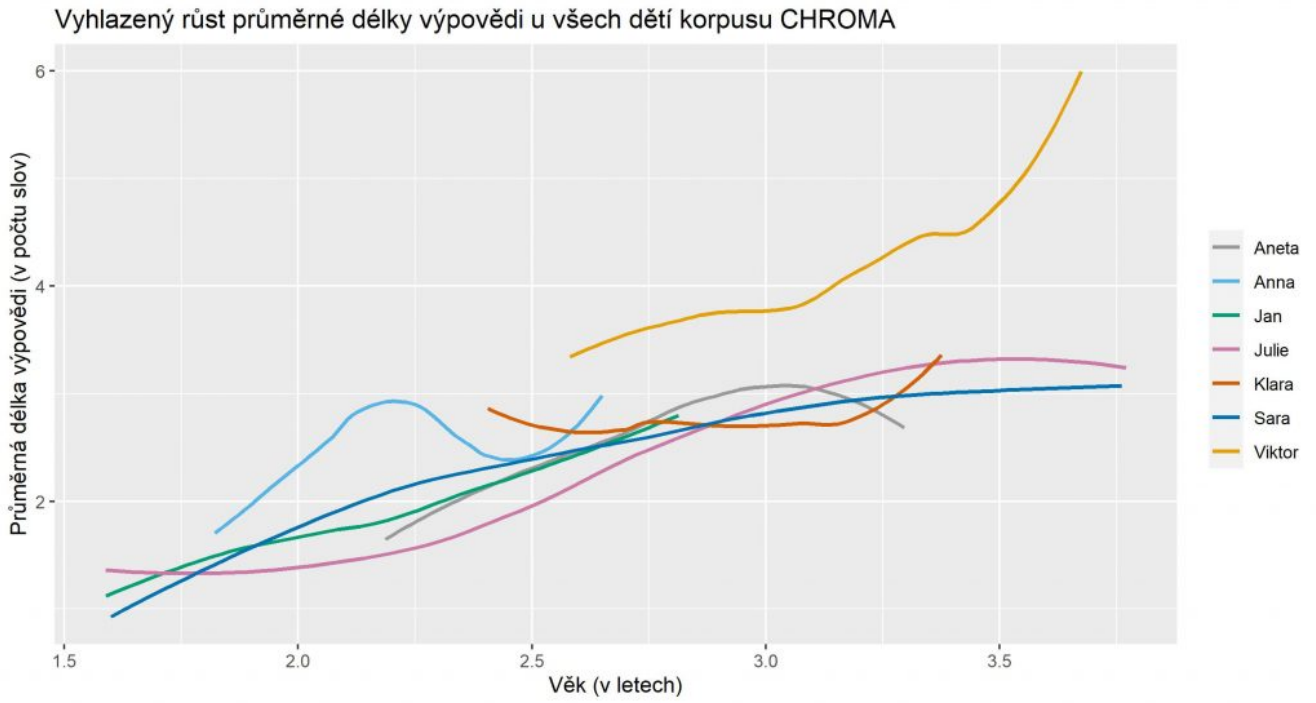
Pro projekt domácího nahrávání ale již nové přispěvatele nehledáme. Pomalu se chýlí ke konci a my začínáme plánovat nový, zaměřený na děti s vývojovou poruchou jazyka.

Jaké je využití podobných dat? Jsou užitečná například v pedagogii, psychologii, nebo i jiných oborech?

Na našich datech lze sledovat nepřeberné množství věcí. Primární využití určitě bude pro základní psycholingvistický výzkum v oblasti typického osvojování prvního jazyka. Poznání

typického komunikačního vývoje v tomto věku je ale v každém případě klíčové i pro předškolní vzdělávání, pro speciální pedagogiku a zejména pro logopedii. Na pozadí popisu typického vývoje teprve můžeme přikročit k výzkumu a aplikacím v oblasti vývoje netypického. Takže na tom vzdálenějším horizontu uplatnění naší práce je možné vidět třeba i diagnostické a terapeutické nástroje pro děti s jazykovými poruchami.

Moje vlastní práce spadá do oblasti základního výzkumu. Aktuálně se intenzivně věnujeme přípravě témat studentských závěrečných prací využívajících naše data a těšíme se, že teď společně se studenty a studentkami budeme schopni z korpusů vytěžit maximum zajímavého. Budeme se věnovat například měřítku průměrné délky výpovědi. Každé dítě začíná s jednoslovnými výpověďmi a délka postupně roste.



V grafu je vidět nejen zřetelný růst průměrné délky výpovědi v čase u všech dětí, ale i značné individuální rozdíly v dynamice toho růstu.

Julie z korpusu *Chroma*: růst délky a komplexity výpovědí

Věk (R;MM.DD)	Průměrná délka výpovědi	Příkladová výpověď z daného přepisu
1;07.05	1,12	*CHI: pípí@i . *MOT: ano , to je ptáček .
2;01.18	1,35	*CHI: Brumínek tuli:@i . *MOT: chceš se tulit s Brumínkem , jo ?
		*CHI: kluk mašličky . *MOT: ty jsme nalepily i tady tomu klukovi do vlasů , no .
2;04.11	1,56	*CHI: podíváme letadlo . *MOT: no , tak se podíváme na letadlo .
2;06.21	1,75	*CHI: v botech můžeme . *MOT: v botech můžeme , ale ne tady přece .
2;09.05	2,81	*CHI: babička taky má pihy .
3;05.24	3,21	*CHI: kdybysem nepoděkuju , tak nic nebudu dostat .

Poznámka: *@i* je kód pro citoslovce.

Tabulka růst ilustruje pomocí příkladů konkrétních promluv z různých vývojových fází jednoho z našich dětí.

Dál budeme analyzovat třeba růst tvarové diverzity, tedy jaké slovesné nebo jmenné tvary si děti osvojují jako první, a jak tvary přibývají, co mají v tomto směru děti společného a jak moc se liší nebo jak přesně jejich vzájemné odlišnosti souvisí s odlišnostmi ve vyjadřování dospělých, kteří o ně pečují. Zrovna pokud jde o tvarosloví neboli morfologii, je čeština bohatým zdrojem materiálu, který může otázky o jeho osvojování osvětlit podstatně lépe než dosud dominantně studovaná angličtina. Na naše data je ale možné aplikovat třeba i metody konverzační nebo jiné interakční analýzy, je možné zkoumat osvojování syntaxe nebo textové anafory a podobně.

Za jistých podmínek můžeme poskytnout i původní nahrávky, které lze využít pro analýzu osvojování fonologie, ale také třeba jako tréninková data pro modely rozpoznávání řeči. Z videonahrávek je možné analyzovat gesta. Některé z těchto zmíněných témat jsou už teď náplní studentských závěrečných prací nebo i větších projektů mimo náš tým.

V jaké fázi je výstavba korpusu v tuto chvíli? Je už teď někde dostupný?

V rámci projektu CoCzeFLA ve skutečnosti nejde o jeden korpus, ale naší vizí je systém korpusů, které se navzájem doplňují. První korpus s názvem *Chroma*, zahrnující prvních sedm dětí, byl publikovaný už v několika verzích – jednak v databázi LINDAT, jednak v CHILDES, což je mezinárodní databáze orientovaná speciálně na dětské korpusy. Tam je dostupná nejnovější verze korpusu.

Druhý korpus *ChroMat* zahrnuje dalších sedm dětí a bude zveřejněn nejspíš v roce 2025. Přepisy pěti dětí jsou už teď v podstatě hotové, ale jedno další se teprve nahrává a ideálně by mělo skončit až za rok, další je v polovině přepisování. Po finalizaci přepisů následuje ještě morfologická anotace.

Pro první korpus jsme nahrávali jenom audio, intervaly byly trochu chaotické a čtyři ze sedmi dětí začaly nahrávat v době, když už mluvily poměrně hodně. Proto jsme plynule navázali druhým korpusem, který tyto nedostatky kompenzuje. Tyto dva korpusy si jsou nicméně velmi podobné. Nakonec bude nejvýhodnější analyzovat všech 14 dětí společně, eventuálně si z nich vybírat ty, které splňují výzkumné požadavky bez ohledu na původní korpus.

Jaké další korpusy máte v plánu vytvořit?

Náš nejbližší další plán je zmiňovaný korpus jazyka dětí s vývojovou poruchou. Trochu vzdálenějším plánem je ještě jeden korpus typicky se vyvíjejících dětí, ale ve starším věku, přibližně od čtyř až do osmi nebo deseti let, protože představa, že osvojování jazyka je ve čtyřech nebo pěti letech zkrátka hotové, není úplně přesná.

Kdo všechno se na budování korpusu podílí?

Veškerou koncepční práci, plánování a organizaci děláme ve dvou společně s kolegyní Klárou Matiasovitsovou, doktorandkou na FF UK. Pro radu si chodíme zejména k našemu školiteli doc. Filipu Smolíkovi, který působí na Psychologickém Ústavu AV ČR a na Ústavu obecné lingvistiky FF UK. Přepisování pro nás

dělají pregraduální studující.

Máme zavedený systém úvodního samoškolení, který nám výrazně usnadňuje práci při hledání motivovaných lidí s talentem pro naslouchání nedokonalé dětské řeči. Studujícím ale nabízíme také možnost absolvovat u nás povinnou praxi. Studující po zapracování mají kromě přepisu na starosti také revidování po sobě navzájem, což je neméně důležitá část práce, a eventuálně také ruční kontroly automatické morfologické anotace.

Morfologická anotace zásadně usnadňuje prohledávání korpusu, díky ní člověk může například vyhledat všechny tvary jednoho slova najednou nebo naopak třeba všechna podstatná jména v akuzativu. Klíčovou roli v rozvoji morfologické anotace korpusu *Chroma* sehrál kolega doktorand české a anglické filologie Jakub Sláma, který byl součástí projektu START kolegyně Kláry Matiasovitsové.

Studující z principu musí s ukončením studia ukončit i spolupráci s námi. Na podzim 2023 jsme na přepis přijali pět nových studujících a je nás teď včetně mě a Kláry deset, což je za historii projektu nejvyšší počet. Velmi doufáme, že někdo ze stávajícího týmu bude motivovaný k doktorskému studiu a připojí se k týmu CoCzeFLA dlouhodoběji i na koncepční úrovni.

Přestože je čeština relativně malý jazyk, existuje v ní spousta regionálních dialektů, které dítě může snadno pochytit. Odrážíte v korpusu nějakým způsobem i tento aspekt češtiny?

Variace v jazyce je pro budování korpusu složité téma. U korpusu se asi obecně předpokládá, že by se měl alespoň blížit tomu, že reprezentuje daný jazyk v celistvosti, což je podle mě velmi zjednodušená představa i u mnohem větších korpusů, než je ten náš.

Naši účastníci žijí nebo žili v době nahrávání převážně v Praze nebo jejím nejbližším okolí, pouze jedna rodina je

z Hradce Králové. Všichni si ale pochopitelně ve svém řečovém projevu nesou nějaké prvky typické pro jiná místa, kde dříve žili. Nikdo z dospělých účastníků tak nepoužívá varietu češtiny, kterou by bylo možné beze zbytku označit jako obecně českou nebo třeba ostravskou. Všichni kombinují různé prvky a jejich děti to po nich samozřejmě opakují.

Mgr. Anna Chromá vystudovala bohemistiku na Univerzitě Pala

ckéh
o.
Na
Filo
zofick
é
fa
ku
lt
ě
UK
ny
ní
st
ud
uj
e
do
kt
or
sk
é
st
ud
iu
m
ob
ec
né
li
ng
vi
st

iky. V roce 2014 iniciovala korpusový projekt Korpusy raných češtiny. Od března 2021

o
Ko
rp
us
y
pe
ču
je
z
po
zi
ce
ku
rá
to
rk
y
da
t
v
rá
mc
i
ve
lk
éh
o
in
fr
as
tr
uk
tu
rn
íh
o
pr

projektu LINDAT/CLARIAH-CZ na FF UK. Je členkou rady Centra pro digitální humanities

FF UK. V disertačním výzkumu a dalších projektech se věnuje raném u osvojování zá

jmenjá/ty a první/druhé slovesné osoby; osvojování teoretického porozumění

ybným přesvědčením; osvojování slovosledu a pádové morfologie v češtině a němč

in
ě.

Centrum pro digital humanities sdružuje členky a členy akademické obce FF UK se zájmem o digital humanities. Posláním centra je podpora a koordinace pedagogické, vědecké, výzkumné a osvětové činnosti na fakultě zaměřené zejména na oblast digital humanities. Jde například o rozvoj metodologie pro získávání, zpracovávání, analýzu, prezentaci a uchovávání dat v digitální podobě, navazování spolupráce s dalšími pracovišti podobného zaměření, vytváření podmínek pro tvorbu a sdílení dat, přípravu grantů a projektů a zprostředkování výsledků širší odborné i laické veřejnosti.

Projekt Lindat/Clariah-CZ je realizován za podpory Ministerstva školství, mládeže a tělovýchovy ČR v rámci velkých výzkumných infrastruktur pod kódem LM2018101.