Podoba a využití korpusu jinojazyčných a romských mluvčích češtiny: CZESL a ROMi
Zuzanna Bedřichová – Karel Šebesta – Kateřina Šormová – Svatava Škodová

1.1 Akviziční korpusy

Důležitost elektronických korpusů pro jazykové vyučování byla rozpoznána už při jejich vzniku. Poměrně záhy začaly být vytvářeny i speciální korpusy s pedagogickým zaměřením. V našem příspěvku si všimneme jednoho jejich typu, totiž korpusů, které primárně slouží zachycení procesu osvojování jazyka, ať už prvního, nebo druhého/cizího.[1]

Oba typy korpusů, osvojování prvního i druhého jazyka, mají řadu rysů společných: (a) používají podobné techniky při sběru a zpracování jazykových dat, (b) postupují podobně při sledování a zaznamenávání metadat, (c) slouží podobným badatelským záměrům, (d) musí se vyrovnávat s podobným problémem nestandardnosti či „chybovosti" jazykového materiálu získaného od mluvčích s komunikační kompetencí, která se teprve vyvíjí, resp. odlišnou od komunikační kompetence dospělého rodilého mluvčího, (e) budují a využívají k tomu účelu podobné nástroje pro zpracování jazykových dat a vyhledávání v nich, především specifický typ chybových anotací (viz samostatný příspěvek). Přesto se o nich jen zřídka pojednává souhrnně a chybí pro ně i jednotné označení.

Ustálený je termín *learner corpus* užívaný pro označení korpusů jazyka druhého, popř. jemu odpovídající domácí protějšek v příslušném jazyce. V češtině se v tomto významu začíná prosazovat termín *korpus žákovský*. Korpusy osvojování prvního jazyka jednotné označení nemají. Bývají označovány opisnými souslovími, jako korpus jazyka mládeže, jazyka dětí apod., popř. jinak motivovanými termíny, které lze v zásadě vztáhnout i ke korpusům žákovským (*developmental corpus*[2]).

Pokládáme za účelné pracovat s termínem jednotným pro oba typy korpusů, a protože jejich primárním účelem je sloužit výzkumu osvojování jazyka, popř. učebním aktivitám s osvojováním jazyka spojeným či k němu směřujícím, volíme pro jejich označení termín akviziční korpusy.

Akviziční korpusy se od běžných korpusů synchronních odlišují řadou znaků. Nejvýznačnější je přirozeně povaha jazykového materiálu, na němž jsou založeny. Zatímco národní synchronní korpusy se zaměřují na reprezentativnost ve vztahu k současnému jazyku, korpusy akviziční zaznamenávají jazyk předškolních dětí, jazyk školní mládeže či mezijazyk nerodilých mluvčích, tedy struktury velmi nestálé, proměnlivé, závislé na řadě vnějších faktorů, které se od běžně užívaného jazyka mohou někdy velmi podstatně lišit.

Z povahy jazykových dat pak plynou specifické rysy další, např. odlišnosti v pojetí autenticity sbíraných textů, odlišný způsob jejich získávání, odlišná pozornost věnovaná zaznamenávání metadat apod.[3]

Podle časového zaměření sběru rozlišujeme akviziční korpusy *průřezové*, které zaznamenávají jazyk dětí/žáků v jediné etapě jejich vývoje, korpusy *longitudinální*, zachycující jazykový vývoj téhož dítěte/žáka nebo skupiny týchž dětí/žáků v několika různých etapách vývoje, a korpusy *pseudolongitudinální*, které zachycují jazyk různých žáků

---

[1] Pomíjíme v zájmu stručnosti rozdíl mezi pojmy jazyk druhý a jazyk cizí; podrobněji k tomu např. Muriel Saille-Troike 2006. Z praktických důvodů budeme užívat termín *druhý jazyk* jako označení univerzální, *jazyk cizí* pouze diferenčně pro odlišení jazyka, který si mluvčí osvojuje mimo společenství, v němž se daný jazyk běžně užívá – pokud je takové odlišení potřebné.

[2] Srov. McEnery, R. Xiao, Y. Tono 2006, 65: *The term **learner corpus** is used here as opposed to a **developmental corpus**, which consists of data produced by children acquiring their first langure (L1). While L2 learner data for longitudinal analysis can also be called „developmental data", we use the term **developmental corpus** specifically for L1 data as opposed to **learner corpus***. (Ztučnění užito místo kurzívy v originále. – K. Š.)

[3] Podrobněji o specifických rysech akvizičních korpusů Šebesta 2010.

v několika různých etapách jejich jazykového vývoje. Časté jsou rovněž *korpusy smíšené*, které kombinují longitudinální a pseudolongitudinální data.

1.2 AKCES/CLAC

S využitím zkušeností z Českého národního korpusu a domácí tradice výzkumu jazyka dětí a mládeže byly přibližně před šesti lety zahájeny práce na rozsáhlém projektu budování akvizičních korpusů češtiny. Název celého souboru procházel určitými proměnami, dnes ho označujeme jako AKCES – Akviziční korpusy češtiny, resp. CLAC – Czech Language Acquisition Corpora. Projekt byl zahájen a v prvních letech i řešen primárně na Ústavu českého jazyka a teorie komunikace ve spolupráci s Ústavem Českého národního korpusu, postupně k němu přistupovaly ústavy a spolupracující instituce další. Dnes jde o poměrně rozsáhlý okruh spolupracujících institucí univerzitních i mimouniverzitních. Součástí AKCES jsou v současné době dva korpusy hotové (SCHOLA 2010 a EDUCO), jeden těsně před dokončením (SKRIPT), dva ve fázi rozpracování (CZESL a ROMi) a dva ve fázi přípravy (korpus přepisů nahrávek jazykových hodin a IUVENT, korpus mluveného jazyka mládeže).

Korpus SCHOLA 2010 obsahuje přepisy nahrávek vyučovacích hodin všech předmětů v různých ročnících školní docházky a je dostupný na stránkách ČNK. Má rozsah 799 300 slov, odpovídající počtu 204 nahraných a přepsaných vyučovacích hodin o celkové délce 143 hod. 25 minut. Korpus EDUCO obsahuje týž materiál, ale v podobě souvislých textů, není veřejně přístupný, ale mohou ho využívat badatelé na základě žádosti a slouží v jisté míře i potřebám výuky. Vybudování obou korpusů je velkou zásluhou koordinátorky H. Goláňové.

Korpus SKRIPT obsahuje přepisy písemných prací českých žáků různých ročníků a typů škol. V současnosti má rozsah přibližně 600 000 slov. V roce 2011 bylo zahájeno jeho chybové a lingvistické značkování, v r. 2012 by měla být zveřejněna první verze. V dalších letech předpokládáme jeho další rozšiřování. Všechny dosud uvedené korpusy zachycovaly v zásadě češtinu rodilých mluvčích.

2.1 Žákovské korpusy

Žákovské korpusy představují ve výzkumu osvojování jazyka i v jazykové didaktice zásadní přínos. Jejich inovativnost spočívá především v tom, že nabízejí velmi rozsáhlé soubory dat pro zkoumání mezijazyka žáků, a to ve vývoji (dávají tedy např. možnost studovat předpokládanou přirozenou posloupnost osvojování jazykových prostředků) a ve vztahu k významným činitelům, o nichž lze předpokládat, že tento vývoj ovlivňují, jako je věk, pohlaví, první jazyk žáka, okolnosti osvojování druhého jazyka, délka formální jazykové výuky apod.

Žákovské korpusy vedly mj. k redefinování obou základních typů analýz užívaných při studiu osvojování druhého jazyka: kontrastivní analýzy a analýzy chybové. Tradiční kontrastivní analýza srovnávala jazyk výchozí a cílový. Žákovský korpus dává možnost postavit do kontrastu mezijazyk – lze ho srovnávat s jazykem cílovým či výchozím (máme-li k dispozici vhodný srovnávací korpus), popř. srovnávat mezijazyky dvou různých skupin žáků. Kontrastivní analýza mezijazyka (CIA) dovoluje odhalit nejen nekorektní odchylky daného mezijazyka od normy jazyka cílového, ale i nadužívání nebo podužívání určitých jazykových prostředků nerodilými mluvčími ve srovnání s rodilými.

Rovněž pro chybovou analýzu přinesly žákovské korpusy zásadní novum, jednak ve využívání systematické chybové anotace (o ní podrobněji v samostatném příspěvku), jednak v tom, že se chybná užití sledují důsledně na pozadí užití korektních, že si můžeme mnohem důsledněji všímat souvislostí chyb s různými externími faktory, sledovat případná funkční využití formálních odchylek od normy, aspekt cizosti apod.

Výsledky těchto analýz se už dlouho využívají při tvorbě jazykových slovníků, učebnic a dalších příruček pro studenty, především v angličtině. Ta v žákovských korpusech zatím naprosto dominuje, jiné jazyky jsou zastoupeny výrazně méně. Celkový rozsah jazykových dat zachycených v žákovských korpusech jiných jazyků zpravidla nedosahuje jednoho milionu slov, zatímco angličtina je zachycena v korpusech o celkovém objemu kolem jednoho sta milionu slov. Ze slovanských jazyků, pokud je nám známo, disponuje žákovským korpusem (o rozsahu 35 000 slov) pouze slovinština.

O významu a bezprostřední využitelnosti žákovských korpusů pro tvorbu slovníků a učebních materiálů svědčí i to, že si velká nakladatelství zaměřená na vydávání této literatury vybudovala velmi rozsáhlé komerční žákovské korpusy angličtiny, např. *Cambridge Learner Corpus* a *Longman Learners' Corpus*. Z korpusů nekomerčních, vzniklých na akademické půdě, je potřeba jmenovat alespoň *International Corpus of Learner English* (ICLE), který vznikal od počátku 90. let v lovaňském centru CECL jako první akademický korpus tohoto typu.[4]

2.2 První žákovský korpus pro češtinu

Budování prvního žákovského korpusu pro češtinu C2J a v jeho rámci i korpusu jazyka romských žáků bylo zahájeno v r. 2009 ve spolupráci Technické univerzity v Liberci a Univerzity Karlovy v Praze v rámci projektu *Inovace vzdělávání v oboru čeština jako druhý jazyk*.[5] Dokončen by měl být počátkem roku 2012, ale práce na něm budou pokračovat i v dalších letech.

V tomto projektu slouží korpus především jako pedagogický prostředek: studenti učitelství se při práci s ním seznámí s možnostmi, které žákovské korpusy pro práci škol nabízejí, a naučí se s nimi pracovat, popř. vytvářet pro vlastní pedagogickou potřebu vlastní korpusy malého rozsahu. Vedle toho předpokládáme jeho rozsáhlé využití badatelské (už nyní je s ním spojeno několik projektů disertačních prací, nemluvě o pracích magisterských a bakalářských) a didaktické, a to jak bezprostřední, tak zprostředkované.

A. Velikost a první jazyk

Korpus je plánován v celkovém rozsahu cca 2 miliony slov, mezi neanglickými korpusy bude tedy patřit k největším, i když část z tohoto objemu bude tvořit srovnávací korpus rodilých mluvčích. Sběr je zaměřen tak, aby byly v korpusu zastoupeny v dostatečném počtu ústní i písemné projevy mluvčích tří jazykových skupin:

(a) jazyků slovanských, tedy blízce příbuzných. Výrazně mezi nimi převažují mluvčí s ruštinou nebo jiným východoslovanským jazykem, významněji jsou zastoupeni rovněž Poláci, další slovanské jazyky jen minimálně;

(b) jiných jazyků indoevropských. V této skupině jsou podle očekávání převážně mluvčí s francouzštinou, němčinou, angličtinou a španělštinou jako prvními jazyky;

---

[4] Podrobnější přehled lze nalézt např. in Šebesta 2010.

[5] Projekt (reg. číslo CZ.1.07/2.2.00/07.0259) se řeší v rámci OP Vzdělávání pro konkurenceschopnost, je financován ze zdrojů Strukturálních fondů EU – Evropského sociálního fondu a ze státního rozpočtu ČR. Příjemcem dotace je Technická univerzita v Liberci, na řešení se jako partneři podílejí Univerzita Karlova v Praze a Asociace učitelů češtiny jako cizího jazyka. UK jako partner zajišťuje především vznik korpusu. Z TUL se na řešení podílí Katedra českého jazyka a literatury FP, z UK kromě Ústavu českého jazyka a teorie komunikace (sběr dat a metadat, přepis, podíl na anotacích) především Ústav bohemistických studií (sběr dat a metadat), Ústav teoretické a komputační lingvistiky (příprava a realizace anotace, včetně anotace chybové) a Ústav jazykové a odborné přípravy (sběr dat a metadat). Kromě toho spolupracují na řešení i četná pracoviště neakademická, především školy a občanská sdružení, a také řada studentů doktorského, magisterského i bakalářského studia obou univerzit. Podrobněji viz www.C2J.cz.

(c) jazyků nepříbuzných; mezi nimi převažují zejména jazyky dálněvýchodní, především čínština a vietnamština, a arabština.

Zvláštní postavení mají v projektu jazyková data romských žáků. U nich nelze jednoznačně určit, zda je pro ně čeština jazyk první, nebo druhý. Sociální, kulturní i jazykové odlišnosti mezi majoritní českou neromskou komunitou a některými komunitami romskými jsou však takového druhu, že lze předpokládat určité specifické rysy jazykového vývoje u romských dětí a také působení některých odlišných mimojazykových faktorů. K charakteristice samostatného korpusu ROMi i jazykového materiálu získaného od romských dětí se vrátíme v druhé polovině studie.

B. Další relevantní parametry

Český žákovský korpus se vyznačuje snahou o pokud možno co největší otevřenost.

a) Zahrnuje projevy písemné i mluvené. Psaná složka z praktických důvodů převažuje. Texty se sbírají v podobě rukopisné a přepisují se tak, aby bylo v přepisu zachováno maximum informací, tedy např. historie textu – změny, které v něm žák při psaní provedl (škrty, vpisky, změna slovosledu či větosledu, záměna výrazu, změna tvaru apod.), zásahy učitele či úseky textu pocházející od jiné osoby. Do první přepsané verze se promítají i nejednoznačnosti rukopisu – přepisovač v něm zaznamenává, pokud je na pochybách, jak určité písmeno nebo výraz číst. To neplatí přirozeně o kvalifikačních pracích, ty se sbírají v podobě elektronické.

První, kompletní verze přepisu zůstává v archivu a není veřejně přístupná; pro anotování a zveřejnění prochází přepis závěrečnou úpravou, která zachová pouze čistý žákův text „poslední ruky“.

b) Český žákovský korpus se snaží pokrýt pokud možno všechny úrovně znalosti jazyka, od začátečnické až po pokročilou, a zaznamenává je klasifikační stupnicí podle SERR. Tím se odlišuje od většiny světových žákovských korpusů, které se obvykle zaměřují pouze na úroveň jedinou, zpravidla pokročilou. Např. ICLE vymezuje úroveň žáků, od nichž sbírá materiál, jako úroveň studentů angličtiny na univerzitě ve třetím a čtvrtém ročníku studia.

Český korpus zachycuje všechny úrovně, ale nesnaží se v tomto směru o vyváženost – ve sběru psaných dat zřetelně převažují mluvčí úrovní B1 a B2. Je to dáno podmínkami sběru: na těchto úrovních je žáků poměrně nejvíce a také jejich produkce je výrazně delší a bohatší než produkce začátečníků.

c) Rovněž po stránce žánrové bude materiál českého žákovského korpusu různorodý. Některé světové žákovské korpusy se omezují pouze na určitý typ textů, zvl. na texty úvahové a argumentativní. Korpus CZESL neklade v tomto směru žádná apriorní omezení.

Převážnou většinu sbíraného materiálu představují eseje psané jako součást oficiální zkoušky, podobně jako je tomu u většiny světových korpusů, komerčních i nekomerčních. Menší díl tvoří kvalifikační práce, bakalářské, magisterské a doktorské; protože jde o texty vytvářené za odlišných podmínek a také jejich sběr má odlišný charakter, budou tvořit oddělený subkorpus.

C. Metadata

Podrobné záznamy o mluvčích, textech a podmínkách jejich vzniku a sběru jsou u žákovských korpusů naprosto nezbytné: detailností těchto záznamů se žákovské korpusy odlišují od běžných korpusů synchronních.

V českém žákovském korpusu se zaznamenávají u mluvčích kromě běžných sociologických údajů (věk, pohlaví) především údaje vztahující se k jejich jazykové biografii – první jazyk, další jazyky, které zná, pobyt v České republice, případné kontakty s češtinou (např. je-li rodina bilingvní), úroveň znalosti češtiny. Za velkou přednost můžeme pokládat

uvádění úrovně podle SERR. Podrobně jsou zachyceny údaje o kontaktu žáka s češtinou, tedy to, jak dlouho a jak intenzivně se žák češtinu učí, podle jakých učebnic apod.

U textu zaznamenáváme téma, žánr, rozsah textu. Podrobně jsou evidovány podmínky vzniku textu, které vyjadřují míru řízenosti jejich tvorby učitelem, tedy podrobnosti zadání (téma zadáno x nezadáno, žánr zadán x nezadán, rozsah zadán x nezadán, čas zadán x nezadán), opora při tvorbě (povaha a rozsah přípravných aktivit, možnost či nemožnost pracovat se slovníkem nebo jinou příručkou) a také okolnosti sběru (psáno nebo namluveno pro korpus, psáno nebo namluveno jako součást oficiální zkoušky apod.).[6]

Český žákovský korpus bude vybaven rovněž specifickou chybovou anotací; podrobněji se o ní pojednává v samostatném příspěvku.

3. ROMi – korpus romských žáků

Korpus romských žáků vzniká souběžně s korpusem nerodilých mluvčích jako součást téhož projektu. Získávaný materiál se však od jazykových dat nerodilých mluvčích liší a v některých parametrech se liší i povaha sběru a sledovaná metadata.

A. Sběr jazykových dat

Sběr materiálů od romských mluvčích probíhá na dvou úrovních – prostřednictvím spolupráce s pražskými i mimopražskými školami různých typů (jde o běžné základní školy, základní školy praktické a speciální) a prostřednictvím individuálních sběračů, jimiž mohou být například romští pedagogičtí asistenti, pracovníci různých nevládních neziskových organizací, kteří spolupracují s Romy, případně studenti středních a vysokých škol, kteří sbírají materiály přímo v terénu.

Na školách probíhá ve spolupráci s učiteli češtiny zejména sběr písemných materiálů, tedy školních slohových prací, jejichž téma vybírá učitel. Ten také ovlivňuje způsob zadání tématu (vybírá vstupní aktivitu před samostatnou prací; rozhoduje, zda půjde o výběr z více témat, či zda je zadáno pouze jedno téma; diskutuje se žáky o základních pojmech v souvislosti s vybraným tématem apod.).

Autory písemných prací jsou jak romští žáci, tak žáci neromští (procentuální zastoupení je přibližně 64 % Romů a 36 % žáků neromských). Zastoupení chlapců a dívek je vyrovnané, 54 % respondentů tvoří chlapci, 46 % dívky.

Z celkového počtu respondentů tvoří zatím 52 % žáci prvního stupně základních škol (nebo odpovídající věková skupina žáků škol praktických a speciálních), 46 % žáci druhého stupně, 1 % žáci středních škol a 1 % žáci SOU nebo nestudující. To odpovídá našemu přednostnímu zaměření na věkové skupiny odpovídající devíti ročníkům základní školy; v tomto věkovém rozmezí je naším cílem přibližně vyrovnané zastoupení všech věkových skupin.

Při dělení podle typu školy připadá 55 % žáků na základní školy, 34 % na školy praktické, 9 % na školy speciální a 2 % žáků jsou ze SOU nebo nepracující.

Většina písemných prací je poměrně malého rozsahu (přibližně 50 slov na text), délka textů je velmi proměnlivá a bývá ovlivněna zejména věkem žáka.

---

[6] Materiál pro žákovské korpusy obecně jen zřídka představují autentické, přirozené jazykové projevy, vzniklé z autentické komunikační potřeby v autentických situacích reálného života. Jde zpravidla o texty uměle elicitované. Do českého žákovského korpusu zařazujeme z elicitovaných textů pouze ten jejich typ, který lze označit jako projevy *klinicky elicitované* (Ellis, Barkhuizen 2005, 23), tedy které byly elicitovány s pozorností zaměřenou na obsah či funkci sdělení, v simulované komunikační situaci, nikoli texty, jejichž tvorba byla řízena s primárním zřetelem k formě sdělení (texty *experimentálně elicitované*, tamtéž). I klinicky elicitované texty se ovšem mohou lišit co do míry řízenosti. Okolnosti, které míru této řízenosti ovlivňují, je proto potřeba detailně zachytit.

B. Sledovaná metadata

Ke každému textu sběrač vyhotovuje anamnestický dotazník a průvodku. Anamnestický dotazník poskytuje informace o autorovi textu. Jde zejména o věk žáka, třídu/ročník, pohlaví, typ školy, znalost romštiny, první jazyk (který jazyk žák pokládá za svůj první) a komunikační prostředí v rodině (který jazyk/jazyky se hlavně užívají v rodině ke komunikaci, příp. zda někdo v rodině mluví romsky) – získávané údaje se tedy v některých bodech liší od metadat uváděných u korpusu nerodilých mluvčích.

Kromě toho elektronické anamnestické dotazníky přinášejí údaje o tom, zda je místo sběru sociálně vyloučenou lokalitou[7] a o jaký typ školy z hlediska sociologického jde (tradiční městská zástavba/sídlištní škola/venkovská škola).

Průvodka poskytuje informace o textu, ty jsou podobného druhu jako u korpusu nerodilých mluvčích, tedy zaměřené na druh textu, téma, žánr a okolnosti tvorby, popř. sběru.

C. Přepis materiálů a jejich primární zpracování

Zpracování materiálů až do jejich poskytnutí anotátorům probíhá ve třech fázích. První představuje oskenování rukopisu nebo získání zvukové nahrávky, jejich opatření evidenčním číslem a propojení s anamnestickým dotazníkem a průvodkou.

V druhé fázi jsou text nebo nahrávka přepsány podle standardních pravidel a opatřeny kódy a komentáři přepisovače. Podobně jako u nerodilých mluvčích i u textů romských žáků se při přepisu zachovává autentická podoba materiálu včetně chyb, oprav a změn, které provedl žák, zásahů učitele apod. Zároveň je text anonymizován – veškeré osobní údaje a jakákoli další data, která by mohla vést k identifikaci mluvčího, jeho školy nebo bydliště, jsou nahrazována zástupnými výrazy. Ve fázi třetí je text upraven do podoby očištěné od všech cizích prvků a předán k anotaci.

4. Nástin charakteristiky chybovosti textů romských žáků v ROMi

Jak již bylo uvedeno výše, jedním z velkých očekávaných přínosů českého žákovského korpusu i korpusu ROMi je probíhající chybová anotace. Pokusíme se stručně naznačit, jakých typických nejčastějších chyb se dopouštějí romští mluvčí češtiny ve věku cca 6–26 let, na základě chybové analýzy jejich písemných textů ještě před započetím anotace, tedy první chybovou sondu. Na psané texty se zaměřujeme nejen proto, že právě ty budou v této fázi anotovány, ale především proto, že u pořízených rozhovorů, které tvoří orální složku ROMi, lze jen velmi obtížně definovat nějakou závaznou normu, a tudíž i chybovost. Pro školní úspěšnost romských žáků, jejíž zvýšení je konečným cílem ROMi, má navíc rozhodující vliv právě podoba písemného projevu.

Při definici „chyby“ vycházíme z normy závazné pro školu a veřejnost, reprezentované normativními příručkami spisovné češtiny a vtělené do praxe základních a středních škol; za chybu považujeme každou nefunkční odchylku od této normy.

Kromě řady chyb obvyklých u všech žáků (romských i neromských), jakými jsou chybějící písmena, chybná písmena, chyby v interpunkci a diakritice apod., se v textech romských žáků setkáváme s řadou jevů, které se v nich vyskytují velmi výrazně. V této fázi práce si netroufáme tvrdit, zda se jedná jednoznačně o chyby typické pouze pro romské žáky, nebo nikoli, nelze však přehlédnout, že právě v textech romských žáků jsou velmi nápadné, resp. objevují se v nich s nápadně vysokou frekvencí. Lze proto uvažovat o tom, zda by bylo

[7] Viz mapa sociálně vyloučených lokalit, http://www.esfcr.cz/mapa/int_CR.html.

možné alespoň některé z nich považovat za projevy působení jazykově specifického prostředí, v němž romské děti vyrůstají. Některé bychom snad mohli – v souladu s výsledky výzkumu Máši Bořkovcové[8] i s výsledky našeho vlastního výzkumu mluveného projevu romských žáků – považovat za etnolektní. Jde ovšem zatím jen o počáteční hypotézy, které budou moci být prověřeny teprve důkladným systematickým výzkumem.

Při hodnocení textů je nutné mít na zřeteli, že řada z nich pochází od žáků praktických a speciálních základních škol, kteří mohou trpět různými poruchami psaní, zejména dysgrafií. Chyby v jejich textech proto mohou mít také individuální příčiny. Také nároky na písemný projev ve školách tohoto typu bývají zpravidla nižší, což je nutné zohledňovat při srovnávací analýze neromských dětí stejného věku či ročníku, avšak jiného typu školy.

Vzhledem k omezenému prostoru se zde nebudeme věnovat chybám běžně zastoupeným v textech všech žáků, ale zaměříme se pouze na nejvýraznější typy chyb žáků romských. Soustředíme se na chyby plynoucí jednak z často problematického způsobu zápisu (problémy rukopisu), jednak na chyby, které vyplývají z těsné blízkosti písemného textu a mluveného projevu (s rysy etnolektu), která je, zdá se, pro texty romských žáků typická. Jak již bylo řečeno, nástin, který přináší tato studie, představuje základní sondu ještě před zahájením anotace; anotace přinese jistě daleko podrobnější a spolehlivější výsledky.

4.1 Problémy rukopisu

A. Nečitelný rukopis

Při chybové analýze narážíme na několik základních obtíží, které celkovou chybovou analýzu značně ztěžují. Prvním z nich je fakt, že neexistuje způsob, jak hodnotit rukopisná specifika každého autora a jak vyřešit nejasnosti v psaní některých písmen, která jsou při nečitelném rukopise těžko rozlišitelná (zejm. písmena malé *t, l, h, k* a *b*, dále dvojice *o* a *a*). Při přepise bylo uplatněno doporučení, aby přepisovači interpretovali chybovost na úrovni rukopisu „očima učitele“, tj. pokud autor textu při psaní prakticky nerozlišuje malé psací *a* a o, nehodnotí přepisovač každý výskyt těchto písmen jako chybový, ale akceptuje ho jako správný. Tento postup je však natolik subjektivní, že většina přepisovačů volila možnost přepisu obou variant (př.: *Tonečky|Tanečky*). Kromě těchto jevů se setkáváme s žáky oblíbeným sloučením písmen *i* a *y* v jedno (pomocí tečky nad *y*), které znemožňuje jednoznačnou interpretaci a jež lze považovat za záměrné (př. *zájmy|zájmi*).

B. Velká a malá písmena

Dalším výrazným rukopisným problémem, který značně ztěžuje přepis a zkresluje chybovou analýzu, je psaní velkých a malých písmen. Na jedné straně není vždy jednoduché odlišit malá písmena od velkých (např. *s* a *S*), ale vyskytují se i texty, které jsou psané směsicí malých a velkých písmen bez ohledu (zdá se) na jejich funkci. Jako příklad můžeme uvést text, který je napsaný střídavě. Podobně problematické jsou i texty, které jsou psány výhradně velkými písmeny.

C. Chybná písmena

Poslední výraznou skupinu chyb plynoucích z rukopisu autora představují chybně napsaná písmena (např. psací *r* se třemi obloučky). Ta jsou při přepise přepsaná správně a odchylka je

[8] Bořkovcová 2006.

zaznamenána v komentáři přepisovače. Jedná se tedy o skupinu chyb, které jsou při přepise eliminovány.

4.2 Přiblížení mluvenému projevu

Výraznou skupinu chyb romských žáků tvoří jevy, které vypovídají o odvození textu z mluveného projevu. Jsou to jednak případy chybějících písmen, která zachycují hlásky, jež často nebývají vyslovovány, dále spodoba znělosti a chybějící kvantity. Je otázkou, zda lze do této kategorie zařadit i chybějící interpunkci a častou neohraničenost syntaktických celků (text je často koncipován jako jedna dlouhá věta/souvětí).

A. Vynechání písmen

Mezi slova s vynechanými písmeny se řadí především pomocné sloveso *jsem* psané jako *sem,* ale typické je také vynechání koncového *l* u příčestí minulého u sloves typu „tisknout“: *rozhod jsem se*.

B. Spodoba znělosti

Častými chybami plynoucími z přiblížení textů mluvenému projevu jsou případy spodoby znělosti, většinou na koncové hranici slov: *o kom budu<in> psát hnet sen to nevěděl*; *spolu si chodíme často zahrád na plácek*. Někdy však dochází naopak ke znělému vyslovení obvykle neznělé hlásky (před neznělou hláskou), které je zachyceno i v písemném projevu: *„kamožka“; „šestnázt*“.

C. Chybějící kvantity

Při srovnání s nahrávkami mluveného projevu se domníváme, že do této kategorie chyb (plynoucích z blízkosti k mluvenému projevu) můžeme zařadit i řadu chybějících kvantit, které jsou vlivem silného přízvuku na předposlední slabice vyslovovány krátce. Patrné je to např. ve slovech *kamarad, kamaradka, kamoška*.

D. Projevy etnolektu

Do této kategorie řadíme další projevy etnolektu,[9] např. některé změny ve výslovnosti, záměny hlásek apod., které se projektují do písemného textu, např. *šusenki*.

E. Hranice slov

Otázkou je, zda do této skupiny zařadit velmi výrazný jev, kterým jsou chyby v hranicích slov. Nejčastěji se jedná o psaní několika slov dohromady, ale může dojít i k jevu opačnému, kdy jsou slova naopak rozdělena. Jako příklad můžeme uvést následující text: *Jámám moc koně ráda a proto se nanich radavozím mam Jedmoho koně svéha. Jmemujese karmela. Barvu má bílou a černé tečky a mě se bílí koně líbí. A taky siráda vyrazí měka se svejma kamožkama. Třeba dokyna.*

Závěr:

[9] Více k definici a projevům romského etnolektu viz Bořkovcová 2006.

K repertoáru dosavadních korpusů češtiny přistupuje v posledních letech jejich nový typ – korpusy akviziční. K významným rysům akvizičních korpusů patří zvláště: (a) specifická povaha jazykových dat, (b) podrobné zaznamenávání metadat o mluvčích, textech a okolnostech jejich tvorby a sběru, (c) specifický typ chybové anotace a vyhledávání. Jejich hlavní funkcí je sloužit studiu procesů osvojování češtiny jako jazyka první i druhého (zvl. frekvenční analýza, kontrastivní analýza mezijazyka, chybová analýza s oporou o korpus), tvorbě jazykových slovníků a učebních, popř. testových materiálů i přímo jako didaktický nástroj v jazykovém vyučování.

Literatura

AIJMER K. (ed.), 2009, *Corpora and Language Teaching*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

BEDŘICHOVÁ, Z., K. ŠORMOVÁ, K. ŠEBESTA, (v tisku), ROMI – první rozsáhlá databanka romského etnolektu češtiny. *Český lid*, 98.

BEHRENS H. (ed.), 2008, *Corpora in Language Acquisition Research*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

BOŘKOVCOVÁ M., 2006, *Romský etnolekt češtiny*. Signeta, Praha.

ČERMÁK, F. a kol. (ed.), 2006, *Korpusová lingvistika: Stav a modelové přístupy*. NLN, Praha.

ČERMÁK, F., J. KRÁLÍK, K. KUČERA, K., 1997, Recepce současné češtiny a reprezentativnost korpusu. *Slovo a slovesnost*, 58, 117–124.

ELLIS, R., G. BARKHUIZEN, 2005, *Analysing Learner Language*. Oxford University Press, Oxford..

GRANGER, S. (ed.), 1998, *Learner English on Computer*. Longman, London and New York.

GRANGER, S., J. HUNG, S. PETCH-TYSON (eds.), 2002, *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

KRÁLÍK, J., 2001, Vyvážení zdrojů reprezentativního korpusu synchronní češtiny SYN2000. *Slovo a slovesnost*, 62, 38–53.

McENERY, T., R. XIAO, R., Y. TONO, 2006, *Corpus-Based Language Studies. An advanced resource book*. Routledge, London/New York.

PRAVEC, N., 2002, A Survey of learner corpora. *ICAME Journal* [online], 2002, č. 26, 81–114. Dostupné z: http://icame.uib.no/ij26/pravec.pdf

SAVILLE-TROIKE, M., 2006, *Introducing Second Language Acquisition*. Cambridge University Press, Cambridge.

SINCLAIR, J., 2004, *How to Use Corpora in Language Teaching*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

SINCLAIR, J., 1996, *EAGLES Preliminary recommendations on Corpus Typology. EAG*--TCWG--CTYP/P. Version of May, 1996.
Dostupné z: http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html

ŠEBESTA, K. (v tisku-a), Akviziční korpusy. In *Minulost, přítomnost a budoucnost v jazyce a v literatuře. Ústí nad Labem 1. –3. 9. 2010*. Ústí nad Labem, PF UJEP.

ŠEBESTA, K. (v tisku-b), Čeština cizinců v korpusu. In *Přednášky z 54. běhu LŠSS.* Filozofická fakulta UK v Praze, Praha.

ŠEBESTA, K., 2010, Korpusy češtiny a osvojování jazyka. *Studie z aplikované lingvistiky/Studies in Applied Linguistics*, 1, č. 2, 11–34.

ŠEBESTA, K.; ŠKODOVÁ, S., (v tisku), Žákovský korpus a jeho využití pro češtinu jako druhý jazyk. In *Sborník z konference 20 let vývoje didaktiky cizích jazyků*. Technická univerzita v Liberci, Liberec.

ŠTINDLOVÁ, B. (v tisku), Žákovský korpus. Budoucnost pro poznávání akvizice cizího jazyka. In *Minulost, přítomnost a budoucnost v jazyce a v literatuře. Ústí nad Labem 1. –3. 9. 2010.* Ústí nad Labem, PF UJEP.

ŠULC, M., 2001, Tematická reprezentativnost korpusů. *Slovo a slovesnost*, 62, č. 1, 53–61.

THOMAS, J., 2005, Using Corpora in Language Teaching and Learning. *Teaching English with Technology, A Journal for Teachers of English.* 2005, č. 6/1. Dostupné z: http://www.iatefl.org.pl/call/j_soft23.htm

TONO, Y., 2003, Learner corpora: design, development and applications. In *Proceedings of the Corpus Linguistics 2003 Conference.* Lancaster : United Kingdom, 800–809. Dostupné z: http://www.scribd.com/doc/8254550/Learner-Corpora

XIAO, R., 2008, Well-known and influential corpora In *Corpus Linguistics. An International Handbook. Eds.* A. Lüdeling, M. Kytö. HSK 29. 1. Vol. 1. Mouton de Gruyter, Berlin/New York, 383–457.