Korpusy češtiny a osvojování jazyka

Karel Šebesta
Ústav českého jazyka a teorie komunikace, FF UK, Praha
karel.sebesta@ff.cuni.cz

Abstract:
The paper is focused on special electronic corpora created roughly during the last two years for the study of language acquisition and for didactic use. In the first part of the paper, the history of the research of language acquisition is described, with respect to methods of collecting and processing of the language data. The second part of the paper is focused on acquisition corpora and introduces the acquisition corpora built for Czech.

Key words:
special corpora, language acquisition, learning Czech

Úvod

Vytváření a užívání korpusů souborů jazykových dat čerpaných z mluvených nebo písemných jazykových projevů dětí nebo jiných osob osvojujících si jazyk, dat získávaných podle jednotné metodiky a zpracovávaných a ukládaných v jednotném formátu s přesnou dokumentací o původu, není v studiu jazykové akvizice věc nová; patří spíše mezi tradiční prostředky studia osvojování jazyka, prvního i druhého, resp. cizího.[1]

Význam sběru verbální produkce dětí/žáků pro studium osvojování jazyka je patrný i z toho, že se k jeho proměnám někdy váže periodizace celého oboru. Když např. D. Ingram (1994, s. 7–31) podává přehled studia dětské řeči od poslední čtvrtiny 19. století do první poloviny 90. let století 20., rozlišuje periodu deníkových studií (1876–1926), studií založených na tzv. velkých souborech (1926–1957) a periodu studií longitudinálních (od r. 1957 do současnosti, tj. do první poloviny 90. let). Uvedené periody, jak je Ingram vymezuje, se pochopitelně od sebe liší nejen sběrem jazykového materiálu, ale i řadou rysů dalších, především dominujícím pojetím jazyka a jeho osvojování, které práci s jazykovými daty v příslušné době ovlivňovaly; způsob sběru a zpracování jazykových dat jsou však vždy rysem velmi charakteristickým.

1 Výzkum osvojování jazyka pracuje v zásadě s trojím typem dat: (a) s performančními daty, která nezahrnují verbální produkci (lze je získat např. zaznamenáváním neverbálních reakcí prokazujících porozumění, měřením reakčního času, testováním úsudku žáků o gramatičnosti vyjádření apod.), (b) s verbální produkcí žáků, (c) se zprávami žáků o vlastním učení a užívání jazyka, založenými na introspekci (Ellis, 2008, s. 912).
Sběr jazykového materiálu a vytváření sbírek (neelektronických jazykových korpusů) se týká druhého typu – verbální produkce žáků. Ta se může významně lišit podle míry spontaneity a volnosti, resp. naopak řízenosti. Na jednom pólu stojí přirozené, volné, neřízené vyjadřování v autentických situacích reálného života, na druhém pólu vyjadřování, které je jasně řízeno s primárním zřetelem k formě sdělení (experimentálně elicitované projevy). V širokém pásmu mezi oběma póly se nacházejí projevy, které byly rovněž elicitovány, ale s pozorností zaměřenou na obsah či funkci sdělení, nikoli na jeho jazykovou formu. Tyto projevy bývají označovány jako projevy klinicky elicitované (Ellis – Barkhuizen, 2005, s. 23). Patří k nim i eseje a slohové práce vytvářené pro školní účely. Míra jejich řízenosti může být různá v závislosti na řadě faktorů, především na přesnosti a způsobu zadání, na povaze činností organizovaných před vlastní verbální produkcí, na povaze, síle a frekvenci intervenčních zásahů učitele nebo výzkumníka při produkci atd.

Pět period rozlišil ve svém stručném přehledu B. MacWhinney (2000): počáteční periodu impresionistických pozorování, od 70. let 19. století etapu dětských biografií založených na deníkových záznamech, od konce 50. let 20. století etapu přepisů magnetofonových nahrávek, od 80. let dvě na sebe těsně navazující etapy spojené s využitím počítačové technologie. MacWhinney tedy nevyčleňuje zvláštní etapu 20. až 50. let, pravděpodobně proto, že jejím výrazným rysem byly sběry nikoli autentického, nýbrž převážně experimentálně získaného materiálu; zato nově uvádí počátek korpusových sbírek od let osmdesátých.

Pokud uvažujeme o sběru a zpracování autentických, resp. klinicky elicitovaných jazykových dat, je zřejmé, že vyčlenění nové etapy spojené s nástupem počítačových technologií a elektronických korpusů je odůvodněné. V samých počátcích byla charakterizována hlavně novou technikou ukládání a dalšího zpracování jazykových dat, k tomu však záhy přistoupily nové rysy další – v metodologii výzkumu, jeho zaměření, v pohledu na povahu jazyka a jeho osvojování.

Česká lingvistika dosáhla v 90. letech minulého století významných výsledků v oblasti korpusů synchronních i diachronních. Speciální korpusy zaměřené na osvojování češtiny však začínají vznikat až v posledních přibližně pěti letech. V předložené studii některé z nich představíme se stručným připomenutím jejich neelektronických předchůdců a analogických korpusů světových.

1. Tradiční typy sběru jazykových dat

První typ sběru autentických jazykových dat představovaly deníkové záznamy, prováděné převážně rodiči, kteří byli většinou zároveň psychology nebo lingvisty. Počátky jejich rozvoje bývají kladeny do 70. let 19. století.[2] Deníkových záznamů se užívalo primárně pro výzkum osvojování jazyka dětmi v nejranější fázi jejich vývoje. Uplatňovaly se i při sledování tzv. pozdějšího jazykového vývoje, tento výzkum však byl rozsahem velmi dlouho podstatně skromnější a zůstával na periferii badatelského zájmu. Soustavně a ve větší míře začal být jazykový vývoj po 6. roce věku sledován až od padesátých let minulého století, zvl. v souvislosti s edukačními potřebami. O skutečném rozvoji jeho studia se pak mluví teprve od konce let osmdesátých.

Totéž lze v zásadě říci i o výzkumu osvojování druhého jazyka. První studie o bilingvismu byly rovněž založeny na deníkových záznamech – srov. častěji uváděné čtyřsvazkové dílo W. Leopolda *Speech development of a bilingual child: a linguist's record* z 30. a 40. let 20. století nebo deníkové záznamy o pokrocích bilingvního dítěte v druhém jazyce u V. Příhody (1963). O soustavném vědeckém zájmu o výzkum osvojování druhého jazyka však můžeme mluvit rovněž až po 2. sv. válce, kdy se začala konstituovat aplikovaná lingvistika, a o vytváření sbírek pro badatelské účely ve větší míře patrně až od let 60. v souvislosti se sběrem materiálu pro potřeby analýzy chybové.

Mezi hlavní přednosti deníkových záznamů patřila dlouhodobost a pravidelnost sledování. Zpravidla šlo o záznamy prováděné po několik let, u prvního jazyka obvykle po dva až čtyři roky, ale někdy i podstatně déle. Citovaná práce Sternových byla (pozn. 20 založena na sběru

2 Této techniky se využívalo nejen při výzkumu osvojování jazyka, ale i jazykových poruch. Z mnoha děl vzniklých na jejím základě je nejznámější práce Clary a Williama Sternových *Die Kindersprache* (1907). Z výzkumů jazykových poruch lze připomenout např. Wernickeovo dílo o afázii *Die Aphasische Symptomenkomplex* (1874).

trvajícím déle než 18 let, její autoři se ovšem zaměřovali hlavně na první dítě a rané fáze jeho jazykového vývoje.

Výhodou bylo rovněž to, že pozorovateli byli pravidelně rodiče, tedy lidé, kteří dítě velmi dobře znali a díky dlouhodobému pozorování, intimnímu kontaktu s ním a znalosti kontextu mohli jeho projevy více či méně spolehlivě interpretovat. Třetím významným rysem je fakt, že v denících jsou zaznamenávány projevy autentické, nikoli experimentálně elicitované, a že díky důvěrnému vztahu rodič–dítě a pravidelnosti záznamů můžeme pokládat za oslabené působení paradoxu pozorovatele, který jinak výsledky podstatně zkresluje.[3]

Už u těchto záznamů se tedy setkáváme se čtyřmi základními kvalitami, s nimiž se různými cestami vyrovnává i dnešní korpusová lingvistika: dlouhodobost a pravidelnost sledování, poměrně vysoká míra autentičnosti materiálu a znalost kontextu.

Značnou nevýhodou byla ovšem subjektivnost záznamů, která se projevovala nejednotnou metodologií, nestálostí podmínek pozorování a záznamů, nejednotností jejich formátu apod., a také velká omezenost vzorku. Pozorování se vázala na jednotlivé děti, jen zřídka na dvojice nebo trojice dětí.[4] Ze známých deníků se na sledování většího počtu dětí zakládají např. záznamy J. N. Baudouina de Courtenay, zůstaly však nepublikovány a nevzešel z nich, pokud je známo, ani žádný jeho výzkum. To je další nevýhoda deníkových záznamů – bývaly vytěženy a zveřejněny jen z malé části a mnohdy zůstávaly nezveřejněné.

Další typický rys, selektivnost záznamů, byl ve své době nepochybně nevýhodou: deníkové záznamy zachycovaly pouze to, co pozorovatelé (rodiče) pokládali za významné, nepodávaly tedy úplný obraz autentické řečové praxe dětí, zvláště ve vyšším věku, nedovolovaly měřit a posuzovat např. relativní frekvenci výskytu některých jevů apod. V dnešní době, kdy máme k dispozici záznamy kompletní, se selektivní záznamy mohou využívat jako užitečný komplement: výhodou je to, že pozorovatel se při nich může soustředit pouze na jevy s nižší frekvencí a vyhodnotit díky tomu jejich užívání s větší přesností. Toho se využívá v tematických deníkových záznamech, které slouží jako doplněk jiných typů sledování.

Snaha odstranit slabiny deníkových záznamů, především vyloučit vliv subjektivních činitelů, vedla od 20. let k získávání materiálu od relativně velkých skupin dětí se stejnými věkovými i sociálními charakteristikami v standardizovaných experimentálních podmínkách. Získané soubory byly sice poměrně velké, jejich velikost se však odvozovala od velkého počtu respondentů. Vzorek od každého jednotlivce nebýval rozsáhlý a jeho velikost byla zpravidla pevně stanovena. S vyloučením subjektivních činitelů a zvětšením velikosti vzorku se bohužel ztratily i zřejmé výhody starších deníkových sběrů, zvláště dlouhodobost sledování a autenticita projevu. Slabou stránkou těchto výzkumů bylo i zaměření na vnější a poměrně snadno měřitelné charakteristiky, jako byla velikost slovníku apod.[5]

Od nástupu nového pojetí jazyka a jeho osvojování v 60. letech se tento přístup jevil jako neuspokojivý a byl nahrazen výzkumy propracovanějšími. D. Ingram je označuje jako longitudinální, H. Behrensová (2008) však oprávněně poukazuje na skutečnost, že longitudinální

3 Paradox pozorovatele je obecný filozofický problém. V souvislosti s metodologií výzkumu jazyka dětí a mládeže o něm s oporou mj. o zkušenosti W. Labova a dalších pojednává např. S. Romaineová (1984, s. 14n).

4 Např. už uváděná práce Sternových čerpala z pozorování dvou dětí, záznamy V. Příhody (1963) se týkaly jediného chlapce atd.

5 Podrobně rozebírá přednosti a slabiny těchto výzkumů D. Ingram (1994, s. 11–21).

byly i deníkové záznamy. U nového způsobu sledování šlo o svého druhu návrat k některým přednostem deníkových záznamů, resp. o kombinaci výhod obou předchozích přístupů: děti byly sledovány opakovaně po delší dobu (několik měsíců, rok či dva) v pravidelných intervalech a v standardních podmínkách; byly vybrány podle určitých kritérií, materiál od nich sbírali vnější pozorovatelé, nikoli rodiče, dbalo se na to, aby byly ve skupině minimálně tři děti. Celkový objem získaného materiálu byl větší než dřív a podstatný byl i rozdíl v badatelském využití záznamů – nezkoumaly se pouze vnější rysy, ale jazyková kompetence dětí, např. pokud jde o derivaci, inflexi, syntax apod.[6]

Od 50. let se ze skromných počátků začíná rozvíjet také studium jazyka mládeže ve školním věku a dospívání. Nebudeme se jím zabývat podrobněji, přehled staršího výzkumu, popř. celkovou charakteristiku pozdějšího jazykového vývoje lze nalézt u S. Romaineové (1984), M. A. Nippoldové (1998), R. A. Bermanové (2004), J. Androutsopoulose a A. Scholze (1998) či P. Schlobinského, G. Kohla a I. Ludewigta (1993).

V zásadě můžeme rozlišit dvě badatelské linie. První představuje zkoumání tzv. přirozeného jazyka mládeže, zaměřené zvláště na jazyk vernakulární, jazyk výrazných subkultur a komunit dětí a mládeže, kupř. z předměstí velkoměst, ze sídlišť se zajímavou sociální či jazykovou skladbou, s vysokým podílem imigrantů apod.[7] Tento výzkum si primárně všímá rysů pro vyjadřování mládeže specifických, v centru pozornosti bývá slovní zásoba (tzv. slang mládeže), popř. výslovnost. Orientuje se hlavně na jazyk mluvený, často ve vazbě na sociální rozrůzněnost mládeže a osvojování sociolingvistických vzorců v době dospívání. Pokud jde o sběr autentického materiálu, probíhal tradičně formou záznamů z přímého pozorování, později formou nahrávek a jejich analýz. V posledních desetiletích se i zde začaly uplatňovat první elektronické korpusy, srov. analýzu vernakulárního jazyka londýnské mládeže s oporou o korpus (Stenström – Andersen – Hasund, 2002).

Druhou badatelskou linii představuje zkoumání, které bývá často motivováno nároky, jež na jazyk žáků klade škola, resp. které posuzuje jazyk žáků z hlediska jejich postupující *gramotnosti* (zvládání písma, osvojování typických prostředků psaného jazyka a postupů písemného vyjadřování, zvládání širšího spektra písemných žánrů) a *intelektuálního, resp. obecněji kognitivního vývoje* (změny v chápání a užívání abstraktních pojmenování, modálních výrazů, obrazných pojmenování, složitějších syntaktických konstrukcí apod.).

O rozvoji tohoto výzkumu ve světovém měřítku se ve větší míře mluví až zhruba od 80. let minulého století.[8] Jednotlivé výzkumy zaměřené na jazyk školní mládeže však mají delší

6 Podrobnější rozbor lze nalézt opět u D. Ingrama (1994, s. 21n). Typickým příkladem jsou práce R. Browna a kol. o raných fázích osvojování prvního jazyka (např. R. Brown: *First Language: The Early Stages*. Cambridge, 1973).

7 Klasické jsou v tomto ohledu studie U.-B. Kotsinasové o jazyce mládeže v stockholmských předměstích, zvl. Rinkeby, a o švédštině přistěhovalců obecně (např. Kotsinas, 1998; 2005), popř. práce I. Androutsopoulose o recipování mezinárodní subkultury hip hopu mládeží v několika evropských zemích, včetně rysů jazykových (Androutsopoulos – Scholz, 2008), studie o jazyce italských writerů (Romiti, 1998) aj. Z domácích prací můžeme připomenout vedle starších studií o přezdívkách školní mládeže (V. Šmilauera, L. Klimeše a dalších) např. studie o jazyce pařanů (Hoffmannová, 1998), writerů (Svobodová – Kavalová, 1999), hiphoperů (Malá, 2008), o jazyce časopisů pro dívky (Čmejrková, 2000; Kadlecová, 2006; 2007), o webových stránkách mladých lidí (Newerkla, 1999) a další.

8 Za jednu z prvních zásadních publikací se označuje sborník, který editovala M. A. Nippoldová (1988). Starší přehled výzkumu podává už citovaná práce S. Romainové (1984), pro jinou než anglosaskou oblast přibližně z téže doby viz práci C. Sandqvistové a U. Telemana (1989).

tradici. Kdybychom sem zařadili i část výzkumů spojených s Bernsteinovou teorií jazykového deficitu a jeho kritikou, se studiem školní komunikace atd., bylo by potřeba jeho počátky posunout alespoň do let padesátých.

Pokud jde o autentický jazykový materiál, na němž byly tyto výzkumy založeny, pro češtinu lze uvést soubory dvojího druhu. Především soubory úplných projevů žáků, mluvených nebo písemných, které vznikly ve školním prostředí, často jako součást výuky. Sem lze přiřadit např. soubor slohových prací žáků základní školy spojený s výzkumem L. Klimeše z 60. let minulého století[9] nebo soubory mluvených a písemných projevů žáků, které v 70. letech shromáždili samostatným sběrem a podrobili analýze M. Kala a M. Benešová (1976; 1989).

Druhou skupinu představovaly archivy chybových vyjádření. Šlo o sbírky úryvků o různé délce, volené tak, aby byl dostatečný kontext pro interpretaci chyby. Většinou nepřekračovaly hranici větných a souvětných celků. Tyto archivy se uplatňovaly jak při studiu osvojování jazyka prvního, tak druhého/cizího.[10]

Tradice uvedených tří typů sběru jazykových dat, tj. (a) longitudinálního sledování menších skupin dětí v standardizovaných podmínkách, (b) sběrů autentického mluveného vyjadřování žáků ve škole i mimoškolním prostředí a (c) sběrů jejich školních projevů písemných, které dosáhly svého vrcholu v letech sedmdesátých, se staly od 80. a 90. let východiskem pro vytváření speciálních elektronických korpusů zaměřených na osvojování jazyka.

2. Akviziční korpusy

Jako speciální označuje J. Sinclair (1996) takové elektronické korpusy, které nenaplňují zcela alespoň jeden z parametrů, jež se u korpusů očekávají, a nemohou tedy sloužit jako pramen popisu normálního užití daného jazyka. Korpusy jazyka dětí a nerodilých mluvčích k nim Sinclair řadí výslovně.[11]

Speciální korpusy, přesněji korpusy osvojování jazyka, se od běžných korpusů synchronních liší mnoha rysy, např. pojetím reprezentativnosti, autentičnosti textů apod. Moderní korpusy, např. Český národní korpus, se opírají o recepci jako kritérium reprezentativnosti (srov. např. Čermák – Králík – Kučera [1997]; Šulc [2001]). Akviziční korpusy vycházejí naproti tomu vždy z produkce, ale často nejde o produkci aktuální, nýbrž očekávanou či plánovanou, cílovou –

9 Vedle dílčích studií publikovaných zejména v časopise Český jazyk a literatura na nich autor založil svou habilitační práci z r. 1966 *Syntaktický vývoj psaného projevu ve 3.–9. ročníku ZDŠ.*

10 Z našeho prostředí si zaslouží zaznamenání rozsáhlý archiv chybových vyjádření žáků základních i středních škol K. Svobody. Svobodův archiv byl zřetelně výběrový, a i když zaznamenával vnější parametry, jako je věk žáků, typ školy, žánr textu apod., autor s nimi při analýze nepracoval. Doklady byly tříděny systémově podle chybně užitého jazykového prostředku a typu chyby. Z toho pak vycházely dílčí studie o chybách lexikálních, chybách v třetí větě v souvětí apod. (Svoboda, 1972; 1973). O tento typ sběrů se opíraly archivy další, např. vývojově koncipovaný archiv syntaktických nedostatků, s nímž pracoval v 70. letech K. Šebesta (1977), žánrově založený archiv K. Dvořáka a další.

11 „A corpus is assumed to have certain characteristics attached, with default values. Unless stated, these characteristics are attributed to anything called a corpus. A corpus which has one or more non-default values for these characteristics is termed a special corpus: its title should specify its deviations from the assumptions... The special corpora are those which do not contribute to a description of the ordinary language, either because they contain a high proportion of unusual features, or their origins are not reliable as records of people behaving normally... Corpora of the language of children, geriatrics, non-native speakers, users of extreme dialects and very specialised areas of communication (like the heraldic blazon or the knitting pattern, or the auctioneer's pattern) should also be designated special corpora because of the unrepresentative nature of the language involved. "

s pomocí korpusu se snažíme zjistit, jak si žáci osvojují dovednost vytvářet texty určitého typu, třebaže v současnosti do repertoáru jejich aktuální produkce nepatří.

Podobně synchronní korpusy pracují pouze s texty autentickými, resp. nesledují a nerozlišují stupeň jejich řízenosti (třebaže řada textů mediálních byla nepochybně vytvořena na konkrétní zadání, jejich podoba byla ovlivněna vedoucím redaktorem, korektorem apod.), zatímco pro korpusy akviziční má evidence míry autentičnosti, resp. stupně a povahy řízenosti mimořádnou důležitost – proto se vylučují texty elicitované experimentálně, ale i u textů klinicky elicitovaných se důsledně zaznamenávají pokud možno všechny faktory, které mohou míru autentičnosti textu snížit (obsah zadání, jeho způsob, přípravné aktivity, možnost používat slovníky a další pomůcky, situace tvorby, osobnost výzkumníka apod.). Podstatnou odlišností je i výrazně vyšší počet dalších sledovaných a zaznamenávaných proměnných, vázaných na pisatele, sběrače, text, okolnosti jeho vzniku apod.

Souhrnné označení pro korpusy reprezentující osvojování jazyka zatím nemáme a neužívá se ani ve světové korpusové lingvistice. V odborné literatuře se setkáváme pouze s ne zcela ustáleným označením *developmental corpus* pro korpus osvojování jazyka prvního[12] a ustáleným *learner corpus* pro korpus osvojování jazyka druhého či cizího.

Jako ekvivalent termínu *learner corpus* se v české odborné literatuře zpočátku objevilo spojení *korpus studijní* (Čermák – Schmiedtová, 2004); nyní se začíná díky budování prvního českého korpusu tohoto typu prosazovat termín *korpus žákovský*, použitý autorem tohoto článku poprvé r. 2009. Korpus osvojování prvního jazyka zatím speciální české označení nemá.

Korpusy osvojování jazyka, prvního i druhého, mají mnoho rysů společných; je proto praktické mít pro ně i společné označení. V této studii budeme užívat termínu akviziční korpus. Je výhodný i proto, že jej lze snadno adaptovat pro označení korpusu prvního i druhého jazyka (*akviziční korpus L1* vs. *akviziční korpus L2*).[13]

Vedle toho je potřeba u obou typů rozlišovat podle časového zaměření korpusy zachycující jazyk žáků v jediné etapě jejich vývoje (*korpusy průřezové*, *příčné*, *transverzální*), zachycující jazykový vývoj téhož žáka nebo skupiny týchž žáků v několika různých etapách vývoje (*korpusy longitudinální*) a zachycující jazyk různých žáků v několika různých etapách jejich jazykového vývoje (*korpusy pseudolongitudinální*), popř. korpusy *smíšené* (tj. takové, které kombinují longitudinální a pseudolongitudinální data; to je případ patrně nejčastější).[14]

12 Označení *developmental corpus* se v tomto významu užívá poměrně často, třebaže sami jeho uživatelé někdy naznačují, že jsou si vědomi jeho neadekvátnosti. Srov. např. McEnery, Xiao a Tono (2006, s. 65): „The term *learner corpus* is used here as opposed to a *developmental corpus*, which consists of data produced by children acquiring their first language (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*. "

13 Variantně budeme užívat i spojení korpus jazyka dětí, žákovský korpus apod.

14 Na okraj poznamenáváme, že v odborné literatuře věnované akvizičním korpusům se můžeme setkat i s jiným chápáním longitudinálnosti. Např. J. A. Belzová a N. Vyatkina (2008) rozlišují longitudinální a vývojovou analýzu a s tím i longitudinální a vývojové korpusy „with regard to the density of observation of learner performance over time. While 'longitudinal' may refer to analyses in which waves of data are elicited at more distant intervals (e.g., at the beginning and end of a semester), we reserve the term 'developmental' for those analyses in which learner performance is documented at close intervals or at all points of production". Jde ovšem pouze o okrajové pojetí, které není v souladu s obvyklým chápáním longitudinálních výzkumů.

Akviziční korpusy prvního jazyka

Korpus CHILDES

Tvorbu akvizičních korpusů otevřel počátkem 80. let minulého století projekt CHILDES (Child Language Data Exchange System) C. Snowové a B. MacWhinneyho. CHILDES dnes zahrnuje řadu dílčích subkorpusů s celkovým objemem téměř 45 milionů slov od dětí mluvících 28 různými jazyky. Rozsahem jde tedy o zdaleka největší známý mluvený korpus nejen akviziční, ale největší mezi mluvenými korpusy vůbec. Je téměř pětkrát větší než druhý největší – korpus mluvené dánštiny (9 milionů slov) a devětkrát větší než třetí největší co do velikosti, mluvená složka Britského národního korpusu o rozsahu 5 milionů slov (MacWhinney, 2008, s. 165–166; podrobnější informace o CHILDES viz webové stránky <http://childes.psy.cmu.edu/manuals/chat.pdf>).

CHILDES nezahrnuje pouze projevy monolingvních dětí, ale i záznamy dětí bilingvních a ve velmi omezené míře i projevy cizinců, tj. nerodilých mluvčích, dětí s jazykovou poruchou, nahrávky vyučovacích hodin apod. Velkou předností celého projektu je cílevědomá snaha o co nejširší mezinárodní zpřístupnění shromážděných dat v jednotném formátu; dbá se na dodržování jednotného standardu přepisů a morfosyntaktického kódování, které umožňují porovnatelnost a co nejširší využití korpusových dat. Projekt databáze CHILDES během let 1984 až 2002 mohl zaznamenat vznik více než 2000 článků využívajících její data a/nebo program (MacWhinney, 2008, s. 166).

Korpus SCHOLA[15]

Sběr materiálu ke korpusu SCHOLA probíhal v letech 2005–2008, od r. 2008 pokračuje jeho další zpracování, od konce r. 2010 bude zpřístupněn veřejnosti v rámci Českého národního korpusu. Jde o korpus přepisů nahrávek vyučovacích hodin na druhém stupni základních škol a na školách středních. Zařazeno je sem i několik vyučovacích hodin zaznamenaných na prvním stupni ZŠ, jde ale o objem minoritní. Korpus zahrnuje všechny projevy, které se ve školní komunikaci objevují, tedy projevy učitelů, žáků i případných třetích osob nebo projevy mediálně zprostředkované, pokud byly při vyučování použity.

Primárním cílem, který projekt sledoval, je vybudování dostatečné materiálové základny pro studium školní komunikace jako prostředí, které žáky podstatným způsobem komunikačně formuje, a zároveň jako první arény veřejného dorozumívání, v níž se aktivně uplatňují. Tým se proto snažil o vyváženost korpusu, především pokud jde o různé typy školních předmětů, u nichž lze předpokládat odlišnosti co do povahy jazykové komunikace (viz níže), ale přihlížel i k dalším relevantním faktorům, jako je věk žáka, typ školy, věk učitele apod.

Sekundárně se počítá s tím, že shromážděný materiál může sloužit pro studium jazyka dětí ve výukové (tedy formální) situaci. Jazykové vstupy žáků je možné vyčlenit do samostatného subkorpusu a s ním pak pracovat jako s akvizičním korpusem smíšené povahy (obsahuje jak data longitudinální, tak transverzální, resp. pseudolongitudinální). Výhodou je přitom to, že má badatel k dispozici dostatečně široký kontext (repliky učitele, promluvy spolužáků apod.).

15 Korpus byl vytvořen na Filozofické fakultě Univerzity Karlovy v Praze, projekt probíhá s podporou výzkumného záměru MSM 0021620825 (*Jazyk jako lidská činnost, její produkt a faktor*); vedoucí projektu K. Šebesta, koordinátorka projektu H. Goláňová.

V podobě korpusu jsou veřejnosti zatím k dispozici pouze přepsaná jazyková data. Vedle toho zůstávají v samostatné databance PEDAGOG uloženy nahrávky; perspektivně se uvažuje o propojení zvukových a přepsaných dat a o vytvoření korpusu multimodálního.

Korpus není určen pouze lingvistům, ale i pedagogům, psychologům a studentům těchto oborů. Počítá se rovněž s jeho přímým uplatněním v přípravě učitelů a v dalším vzdělávání pedagogických pracovníků. Byl proto od počátku budován jako korpus veřejně přístupný a perspektivně se počítá i s jeho začleněním – v modifikované podobě – do rámce CHILDES.

Velikost databáze byla stanovena podle velikosti mluvených korpusů PMK a BMK v době, kdy sběr začínal, na přibližný celkový rozsah 700 000 slov. V nynější podobě korpus obsahuje jazyková data v rozsahu 799 300 slov, odpovídající délce nahrávek o rozsahu 143 hodin a 25 minut a počtu 204 nahraných a přepsaných vyučovacích hodin. Jde o rozsah i v mezinárodním srovnání poměrně značný. Na jazyky zastoupené v CHILDES připadá v průměru přibližně půl druhého milionu slov, zastoupení jednotlivých jazyků však zdaleka není proporční; nahrávky vyučovacích hodin a jejich přepisy takového rozsahu zatím k dispozici, pokud je nám známo, nejsou.

Sběr materiálu byl veden tak, aby byly v databázi zastoupeny všechny vyučovací předměty s výjimkou cizích jazyků (komunikace v cizojazyčné třídě má velmi specifický charakter) a tělesné výchovy. Aby mohl být korpus vyvážen, byly jednotlivé předměty sdruženy do skupin podle své předpokládané komunikační blízkosti. Vznikly tak tři větší skupiny:

a) matematika a přírodovědné předměty,
b) předměty společenskovědní a výchovné,
c) předměty technické povahy, profesní předměty a pracovní vyučování.
d) Do samostatné čtvrté skupiny byl zařazen předmět český jazyk a literatura, zvláště proto, že je zaměřen přímo na jazykovou komunikaci, a lze u něj tedy předpokládat určitá specifika.

Logika tohoto rozdělení může být ověřena analýzou, která zatím na rozsáhlejším materiálu provedena nebyla. Na základě předběžných sond však bylo možno konstatovat některé odlišnosti mezi vymezenými skupinami, týkající se např. zastoupení odborné terminologie, práce s vzorci a schématy, poměru věcného a subjektivně zaujatého sdělování, míry vázanosti komunikace na kontext, poměru stylu vyprávěcího a výkladového apod.

V korpusu jsou dvě skupiny, totiž matematika a přírodovědné předměty a český jazyk a literatura, zastoupeny přibližně po 30 %, další dvě skupiny, tj. předměty společenskovědní a výchovné a předměty technické povahy a pracovní vyučování, cca po 20 % (měřeno počtem textových slov; podrobněji viz tabulka). To se proporčně zhruba blíží jejich zastoupení ve vzdělávacích programech, které měl tým k dispozici a z nichž vycházel – vzhledem k tomu, že školy mají při vytváření školních vzdělávacích programů značnou volnost, lze se ovšem při určování proporcí jednotlivých skupin předmětů školní praxi pouze přibližovat.

Zastoupení typů předmětu v korpusu SCHOLA

Typ předmětu	Počet nahrávek	Počet slov	Procento z počtu slov
matematika a přírodovědné předměty	63	268 074	31 %
český jazyk a literatura	60	271 477	29 %
společenskovědní a výchovné předměty	44	169 476	22 %
informatika, technické a profesní předměty, pracovní vyučování	37	161 738	18 %
CELKEM	204	799 300	100 %

Sledované parametry

Všechny sledované a zaznamenávané parametry můžeme sdružit do čtyř skupin. Jsou to parametry: (a) žáka, (b) učitele, (c) předmětu a vyučovací hodiny, (d) školy a třídy.

Sledované parametry v korpusu SCHOLA

Parametry žáka	Parametry učitele	Parametry předmětu a vyučovací hodiny	Parametry školy a třídy
Pohlaví	Pohlaví	Předmět	Sídlo[16]
Věk	Věk	Skupina předmětů	Velikost sídla

16 Hodnoty: Praha, mimopražské sídlo.
17 Hodnoty: i = do 35 let, v = nad 35 let.
18 Hodnoty: převážně jazyková, převážně slohově-komunikační, převážně literární.
19 Hodnoty: tradiční městská zástavba, sídliště, venkov.
20 Hodnoty: do 11 let, 11–15 let, nad 15 let.
21 Hodnoty: do 30 let, 31–40 let, 41–50 let, 51–60 let, nad 60 let.
22 Hodnoty: Čechy, Morava, Slezsko.
23 Hodnoty stanoveny v souladu s členěním nářečních oblastí v Českém jazykovém atlase.
24 Žáci jsou řazeni podle typu započatého studia: studenti středních škol jsou tedy vedeni v kategorii SŠ.
25 Hodnoty: základní škola, střední odborná škola, čtyřleté gymnázium, víceleté gymnázium.
26 Hodnoty: s poruchou učení, s poruchou chování, cizinec, jiné (s komentářem).
27 Hodnoty: škola státní, církevní, soukromá.
28 Hodnoty: převážně výkladová, procvičovací, opakovací, prověřovací, smíšená, jiná.
29 Hodnoty: standardní, smíšená (ž. z několika ročníků), sloučená (ž. z několika tříd téhož ročníku).

Věková kategorie dle ORAL 2006[17]	Věková kategorie dle ORAL 2006	Zaměření ČJL[18]	Typ sídliště[19]
Věková skupina: žák[20]	Věková skupina: učitel[21]	Téma hodiny	Region[22]
Vzdělání (VŠ x ZŠ, SŠ)	Vzdělání (VŠ x ZŠ, SŠ)	Pořadí v rozvrhu	Nářeční oblast[23]
Vzdělání – kategorie podle korp. ORAL[24]	Vzdělání – kategorie podle korp. ORAL	Skupina předmětů	Typ školy[25]
Zvláštní poznámky[26]	Aprobace	Typ (klasická, jiná)	Typ podle zřizovatele[27]
Pobyt v dětství: stát	Pobyt v dětství: stát	Typ podle funkce[28]	Třída/ročník
Pobyt v dětství: nářeční oblast	Pobyt v dětství: nářeční oblast	Místo konání	Typ třídy[29]
Současný pobyt: nářeční oblast	Současný pobyt: nářeční oblast		Počet žáků[30]
	Ostatní – médium		

Korpus SKRIPT[31]

Druhý korpus, korpus písemných projevů žáků ZŠ a SŠ, navazuje na tradici sběru a analýz školních slohových prací. Je založen na přepisech písemných prací žáků od 5. třídy ZŠ po nejvyšší ročníky gymnázií.

Sběr materiálu probíhal v letech 2006–2009, do konce r. 2010 pokračují kontroly přepisů a další úpravy, v r. 2011 je plánováno vytvoření korpusu a jeho zpřístupnění v rámci ČNK. V dalších letech se plánuje jeho další rozšíření a zejména doplnění lingvistických a chybových anotací.

Primárním cílem vytvoření korpusu SKRIPT bylo vybudování materiálové základny pro výzkum vývoje písemného vyjadřování dětí ve věku školní docházky. Sekundárně počítáme s tím, že bude sloužit rovněž jako srovnávací korpus při analýzách písemných projevů žáků s jiným než českým mateřským jazykem a při analýzách jazyka žáků se sociokulturním handicapem.

Plánovaná velikost korpusu v první fázi je cca 560–600 000 slov, s tím, že se korpus bude průběžně dále rozšiřovat, především tak, aby byly pokryty další regiony a také typy škol. Zejména se uvažuje o zařazení prací ze středních škol odborných a učňovských zařízení.

Korpus SKRIPT je vyvažován na základě věku žáků jako hlavního kritéria; k dalším parametrům se přihlíží a jsou důsledně zaznamenávány, aby bylo možné na jejich základě či na základě jejich kombinací vytvářet samostatné subkorpusy pro potřeby konkrétních výzkumů.

Skládá se ze dvou částí. První obsahuje písemné práce žáků druhého stupně základní školy a nižších ročníků víceletých gymnázií. Z každé třídy, resp. ročníku, jsou do něj zařazeny práce v celkovém rozsahu cca 20 000 slov. Druhá obsahuje písemné práce žáků čtyřletých gymnázií

30 Hodnoty: do 20 žáků, nad 20 žáků.

31 Korpus vzniká rovněž na Filozofické fakultě Univerzity Karlovy v Praze, vedoucí projektu K. Šebesta, koordinátorka J. Letafková; projekt probíhá s podporou výzkumného záměru MSM 0021620825 (*Jazyk jako lidská činnost, její produkt a faktor*).

a vyšších ročníků gymnázií víceletých. Z každého ročníku jsou do něj zařazeny práce v celkovém rozsahu cca 120 000 slov. Práce byly získávány zatím převážně ze škol středních a jižních Čech, další regiony budou pokryty v následující etapě rozšiřování korpusu v příštích letech.

Sledované parametry

Při sběru a přepisu textů byla sledována vedle věku žáků řada dalších parametrů, které lze sdružit do čtyř skupin: parametry (a) místa vzniku (regionu, obce, školy a třídy), (b) textu a situace jeho vzniku, (c) žáka, (d) učitele.[32]

Parametry sledované u písemných prací v korpusu SKRIPT

Parametry regionu, obce, školy, třídy	Parametry textu a situace jeho vzniku	Parametry žáka	Parametry učitele
Region	Téma	Pohlaví	Pohlaví
Nářeční oblast	Téma podle zadání[33]	Věk	Věk
Obec	Žánr	Věková skupina: žák	Věková skupina: učitel
Velikost obce[34]	Žánr podle zadání[35]	Pobyt v dětství: stát	Vzdělání
Typ školy	Osnova podle zadání[36]	Pobyt v dětství: nářeční oblast	
Typ podle zřizovatele	Text s osnovou: ANO x NE	Současný pobyt: nářeční oblast	
Třída/ročník	Typ práce: školní x domácí		
	Závažnost práce[37]		
	Rok vzniku		
	Čas na vypracování[38]		
	Způsob zadání, situace psaní[39]		
	Hodnocení práce učitelem		

32 Parametry shodné s parametry u korpusu SCHOLA blíže nekomentujeme.

Akviziční korpusy L2 (žákovské korpusy)

První veřejný korpus jazyka nerodilých mluvčích začal vznikat od r. 1990 v belgické Lovani v Centre for English Corpus Linguistics (CECL) zásluhou S. Grangerové a jejího týmu. Lovaňský korpus ICLE (International Corpus of Learner English), vytvářený pod jejím vedením, zahrnuje argumentativní eseje o rozsahu 500–1000 slov na obecná nebo literární témata od studentů angličtiny jako cizího jazyka ve třetím nebo čtvrtém ročníku. Původně byly tyto eseje získány z 11 zemí s různými prvními jazyky (včetně češtiny); v druhé zveřejněné verzi bylo doplněno dalších 5 zemí, především neevropských, a celkový počet slov se navýšil z 2,5 milionu na 3,5 milionu slov, počet esejů z 3640 na 6085 textů.

Lovaň zůstává hlavním světovým centrem práce s korpusy tohoto typu, dnes zejména vedle čínských a japonských univerzit. Od 90. let se budování žákovských korpusů velmi rychle rozšířilo, mj. díky velkým možnostem jejich využití nejen ve sféře akademické a pedagogické, ale i komerční – pro vytváření kvalitnějších slovníků, mluvnic, učebnic a dalších materiálů určených pro zájemce o studium příslušného jazyka. Proto také vytvářejí své interní, neveřejné žákovské korpusy velká specializovaná nakladatelství. Z nakladatelských korpusů jsou známé zejména Cambridge Learner Corpus a Longman Learners' Corpus, rozsahem podstatně převyšující ICLE.

Ve světě žákovských korpusů vládne dnes velká variabilita, podstatně větší než u akvizičních korpusů prvního jazyka. Není to dáno jen tím, že se u korpusů žákovských nevytvořilo centrum podobné CHILDES, ani větší komerční využitelností žákovských korpusů. Svou roli sehrává rovněž podstatně větší různorodost materiálu, který žákovské korpusy zpracovávají, a také vyšší náročnost na jeho získávání a další zpracování.

Korpusy druhého jazyka můžeme třídit podle různých kritérií: podle cílového jazyka, prvního jazyka žáků, média (korpusy psané vs. mluvené), komerčního vs. akademického určení, podle proměnných, které zaznamenávají, podle úrovně pokročilosti sledovaných žáků, podle objemu jazykových dat, podle vybavenosti anotacemi, zvl. chybovými, apod.

Uvedeme pouze některé relevantní rysy světových žákovských korpusů v úzkém výběru. V souhrnných číselných údajích se opíráme o přehled žákovských korpusů na webových stránkách CECL. Podrobnější informace o tomto typu korpusů lze nalézt v pracích N. Pravece (2002), N. Nesselhaufové (2005), přehled bádání podává S. Grangerová (1998; 2009) a další.

33 Hodnoty: volné, zadané, zadané s osnovou, výběr z určených témat.

34 Hodnoty: do 5000 obyvatel, nad 5000, nad 10 000, nad 50 000, nad 100 000 obyvatel.

35 Hodnoty: volný, zadaný, výběr z určených žánrů.

36 Hodnoty: zadaná učitelem vs. nezadaná učitelem.

37 Hodnoty: vysoká vs. střední vs. nízká.

38 Hodnoty: vyučovací hodina, dvě hodiny, část hodiny, libovolný čas (u práce domácí), jiný čas.

39 Hodnoty: společný výběr tématu, společná inspirace, společná příprava osnovy apod.

Cílový jazyk a první jazyky

Téměř všechny známé světové korpusy L2 jsou zaměřeny na jediný cílový jazyk. Jednou z výjimek je např. multilingvní žákovský korpus budovaný na katolické univerzitě v Sao Paulu, USP Multilingual Learner Corpus (MLC), který obsahuje projevy brazilských studentů v němčině, angličtině a španělštině (Tagnin, 2010).

U naprosté většiny žákovských korpusů je tímto cílovým jazykem angličtina. Z více než 90 korpusů uváděných v přehledu na stránkách CECL je jich angličtině jako cílovému jazyku věnováno cca 60. Převaha angličtiny je ještě patrnější, pokud přihlédneme k uváděnému objemu dat. Z celkového objemu cca 100 milionů textových slov, který můžeme v citovaném přehledu najít, je více než 95 % obsaženo v korpusech angličtiny. Z ostatních jazyků je více zastoupena pouze němčina, francouzština, španělština, švédština a italština, většinou cca půlmilionovými korpusy. Ze slovanských jazyků se uvádí pouze korpus slovinský o rozsahu 35 000 slov.[40]

Pokud jde o první či výchozí jazyky, je jejich repertoár poněkud pestřejší, především proto, že některé korpusy, zvl. lovaňský a rovněž oba uvedené korpusy komerční, zahrnují jazyková data studentů s prvními jazyky různými. Jednoznačnou převahu však mají mluvčí jazyků dálněvýchodních – čínštiny (téměř třetina dat v dosud známých žákovských korpusech pochází od čínských mluvčích), s velkým odstupem japonštiny a korejštiny. Z indoevropských jazyků je kromě angličtiny více zastoupena především němčina a švédština (oba jazyky kolem 1,5 až 2 milionů slov). Čínština jako první jazyk má převahu hlavně díky velkému korpusu angličtiny čínských studentů HKUST (Hong Kong University of Science and Technology Learner Corpus), který obsahuje 25 milionů slov, a řadě dalších, méně objemných korpusů (SWECCL, The Spoken and Written English Corpus of Chinese Learners, cca 2 miliony slov; CLEC, Chinese Learner English Corpus, cca 1 milion slov; MSEE, Corpus for Middle School English Education, 2,3 milionu slov; TLCE, The Taiwanese Corpus of Learner English, cca 2 miliony slov atd.).

Čeština je zatím v žákovských korpusech zastoupena pouze jako první, výchozí jazyk, a to v jednom ze subkorpusů obsažených v lovaňském ICLE s angličtinou jako cílovým jazykem v rozsahu cca 200 000 slov.

Velikost žákovských korpusů

Co do objemu dat žákovské korpusy poměrně silně kolísají. Některé z největších a nejmenších světových korpusů uvádíme v následujících tabulkách. Celkově lze říci (viz též výše), že obsáhlejší korpusy o milionových počtech slov existují v současnosti pouze pro angličtinu jako cílový jazyk, korpusy všech ostatních jazyků se pohybují v řádu statisíců nebo dokonce desetitisíců slov.

40 Uváděné číselné údaje jsou bohužel pouze velmi přibližné a nepřesné. Citovaný přehled žákovských korpusů na webových stránkách CECL vychází z dobrovolných oznámení jejich tvůrců, nemusí tedy zdaleka být úplný a spolehlivý; u některých projektů není vůbec uvedena velikost korpusu, popř. další podstatné parametry. Vzhledem k velké dynamice akvizičních korpusů obecně a zejména korpusů žákovských lze nadto předpokládat, že je v současnosti řada korpusů ve fázi vzniku, ale dosud o nich nemáme zprávy. Přesto mohou myslím tyto údaje posloužit jako určitá nápověď.

Největší známé světové žákovské korpusy[41]

Název	První jazyk	Cílový jazyk	Počet slov
CLC – The Cambridge Learner Corpus	různé jazyky	angličtina	30 milionů
HKUST – The Hong Kong University of Science & Technology Corpus of Learner English	čínština	angličtina	25 milionů
LLC – The Longman Learners' Corpus	různé jazyky	angličtina	10 milionů
ICLE – The International Corpus of Learner English	různé jazyky	angličtina	3,5 milionu

Nejmenší známé světové žákovské korpusy

Název	První jazyk	Cílový jazyk	Počet slov
The Pilot Arabic Learner Corpus	angličtina	arabština	9 000
The Korean learner corpus	různé jazyky	korejština	10 000
Ursula Weinberger's learner corpus	angličtina	němčina	27 635
The EVA Corpus of Norwegian School English	norština	angličtina	35 000
The PIKUST pilot learner corpus	různé jazyky	slovinština	35 000

Získávání materiálu, úroveň mluvčích a typy textů

Materiál pro žákovský korpus se získává poměrně obtížně. Nerodilí mluvčí užívají cílového jazyka (zvláště jde-li o jazyk cizí) převážně pouze ve škole, při zkouškách, výjimečně při jiných příležitostech, popř. od nich lze materiál získávat samostatně jen pro potřeby korpusu, opět ovšem na základě elicitace.

Naprostá většina jazykových dat ve světových korpusech je proto založena na materiálu získaném při zkouškách nebo ve školních podmínkách. To platí jak o korpusech psaných, tak mluvených – mluvený korpus angličtiny jako cizího jazyka SST (Standard Speaking Test) o rozsahu 1 milion slov byl vytvořen z interview a dialogů v rámci standardní zkoušky japonských studentů. Samostatně byla elicitována interview a popisy obrázků pro vytvoření lovaňského korpusu mluvené angličtiny LINDSEI (Louvain International Database of Spoken English Interlanguage). Projevy spontánní v běžných životních situacích se v žákovských korpusech objevují jen velmi zřídka.

Pokud jde o úroveň sledovaných žáků, ne všechny korpusy tento údaj uvádějí; z těch, u nichž uveden je, se naprostá většina orientuje na studenty pokročilé, popř. středně pokročilé, na

41 Některé z uváděných korpusů jsou koncipovány jako korpusy otevřené, jejich velikost tedy stále roste.

studenty cílového jazyka na univerzitách, popř. (u angličtiny) na absolventy středních škol, uchazeče o studium na univerzitách apod. Souvisí to mj. s postojem k jazyku začátečníků – má se za to, že jejich projevy jsou natolik chybové, že se příliš nehodí pro korpusové zpracování.

Např. zmiňovaný ICLE sbírá materiál od studentů angličtiny jako cizího jazyka ve třetím nebo čtvrtém ročníku univerzitního studia. Úroveň začátečník až středně pokročilý či pokročilý je uvedena pouze u cca devíti korpusů menšího rozsahu.

Rovněž typy textů představují repertoár relativně omezený. Z projektů, které tento údaj uvádějí, se více než polovina opírá o eseje, resp. texty ze zkoušek, zkušební interview apod.; zařazovány jsou však i zprávy o výzkumech, disertační, magisterské a seminární práce a další texty vznikající ve standardních školních situacích. ICLE žánrový repertoár zužuje na eseje argumentativní (nikoli třeba texty vyprávěcí). Žánrové omezení přirozeně omezuje reprezentativnost existujících korpusů a platnost výzkumných závěrů na nich založených.

Zaznamenávané parametry

Zaznamenávání různých parametrů o žácích a o způsobu získání psaného nebo mluveného materiálu je pro kvalitu korpusu a jeho využitelnost klíčový. Jak už bylo uvedeno, žákovské korpusy zpravidla sledují podstatně bohatší repertoár různých extralingválních rysů než běžné korpusy synchronní; jejich získávání je však poměrně náročné a ne vždy plně spolehlivé.

Korpus ICLE zaznamenává dvě sady proměnných: proměnné vázané k žákovi a proměnné vázané k úkolu, resp. vzniku textu. Programově si tvůrci ICLE vytkli větší počet sledovaných proměnných (Granger, 1998), ne všechny však byly ve všech národních subkorpusech sledovány. Počet obecně sdílených proměnných je tedy omezenější, uvádíme je v tabulce podle R. Ellise a G. Barkhuizena (2005, s. 339), viz též S. Grangerová (1998, s. 9).

Proměnné sledované ve všech subcorpusech ICLE

Proměnné vázané k žákovi	Proměnné vázané k vzniku textu
Věk	Médium (psaný vs. mluvený projev)
Kontext učení	Obor (obecný, literatura apod.)
Úroveň ovládání jazyka	Žánr
	Délka

Dosavadní výzkumy všech možností, které badatelům zaznamenávání proměnných nabízí, zatím plně nevyužívají a omezují se zpravidla na sledování vlivu prvního jazyka na osvojování jazyka druhého/cizího.[42] Soustředění na první jazyk plyne z převažujícího a pochopitelného

42 V kritickém ohlédnutí na to upozorňuje S. Grangerová (2009, s. 17): „One must admit, however, that this facility is still seldom used and LC researchers (myself included) have had a tendency to base their analysis on the whole corpus or on subcorpora distinguished only on the basis of the learners' mother tongue. In fact, a properly coded learner corpus makes it possible for researchers to study the effect of a much wider range of variables. "

zájmu o interferenční jevy, lze ale očekávat, že s rozvíjejícím se korpusovým výzkumem v této oblasti začnou badatelé těžit i z dalších možností, které detailní záznamy o jazykových datech v žákovských korpusech nabízejí.

Chybové značkování

Žákovské korpusy procházejí obvyklým lingvistickým značkováním. Jde však o věc značně náročnou, zejména pokud jde o texty vysoce chybové – v takovém případě nelze zpravidla využít značkování automatické a je nutné věnovat se časově i finančně náročnému značkování manuálnímu.

Specifickým typem značkování žákovských korpusů, který ovšem může být využit pro akviziční korpusy obecně (počítá se s tím např. u korpusu SKRIPT i korpusu SCHOLA), je značkování chybové. Navazuje na tradici někdejší chybové analýzy, ale odstraňuje její slabiny, pro něž byla kritizována a opuštěna.

Především se mění taxonomie chyb tak, aby byla založena pouze na pozorovatelných datech a zahrnovala kategorie, které se nepřekrývají. Podstatné je i to, že chybová analýza s oporou o korpus se uplatňuje v kontextu komplexní analýzy celistvého systematicky shromážděného jazykového materiálu s co nejúplněji dokumentovanými lingvistickými a extralingvistickými daty. Chybově anotovaný korpus se tedy zásadně odlišuje od někdejších heterogenních sbírek chybových vyjádření a s tím i chybová analýza na něm založená od chybové analýzy tradiční.

Chybové značkování však přináší řadu problémů. Jednou z nejvážnějších komplikací je fakt, že chybu lze opravit a v souvislosti s tím i klasifikovat, popsat a hodnotit často různými způsoby, tedy že můžeme u mnoha chyb uvést několik možných hypotéz.

S. Grangerová uvádí, že systém chybových anotací by měl být konzistentní, informativní, flexibilní a využitelný pro různé jazyky. Každý z uvedených požadavků klade na systémy anotací značné nároky. Zatím v této oblasti nebyly vyvinuty mezinárodní standardy a chybová anotace se liší od korpusu ke korpusu.

Chybová anotace je také jen poměrně zřídka systematická/plošná, tj. zřídka se snaží o zachycení celého repertoáru chyb. Mnohem častěji se s ohledem na náročnost anotací volí anotace částečná, tj. anotují se pouze některé chybové jevy podle zaměření konkrétního výzkumu (např. chyby ve flexi sloves) a jiné se nechávají stranou.

Ojediněle se uplatňuje i anotace implicitní, kdy se chyba pouze identifikuje a opraví, ale není k ní připojena žádná chybová značka. Tak postupuje např. korpus MELD (The Montclair Electronic Language Database). Výhodou implicitní anotace je mj. to, že vnáší do značkování méně lingvistické teorie.

Značné rozdíly jsou i ve formátu značkování. Nejčastěji se volí značkování rovinné, tj. chybová značka, popř. i navržená oprava, jsou uvedeny přímo v přepsaném textu před chybou nebo za ní. Zatím řídké je značkování vícerovinné, které umožňuje lépe zachytit několik různých hypotéz opravy chyby. Pracuje s ním např. německý korpus FALKO – Fehlerannotiertes Lernerkorpus (Lüdeling et al., 2006).

Dosavadní systémy chybového značkování byly vyvinuty pro zpracování jazyků s chudou flexí. Pro jazyky s flexí bohatou, jako je čeština, je potřeba vypracovat systémy složitější. Přehled systémů chybového značkování ve světových korpusech podávají např. A. Díaz-Negrillo a J. Fernández-Domínguez (2006).

Obecně platí, že žákovské korpusy jsou stále ještě novou větví v korpusové lingvistice a z toho plynou i některá jejich omezení. K těm nejvážnějším patří nerovnoměrné zachycení cílových jazyků, které nedovoluje rozsáhlejší mezijazykové srovnávání, poměrně omezený repertoár prvních jazyků (většina korpusů je založena na projevech mluvčích s jediným prvním jazykem), nevyrovnané zastoupení různých úrovní ovládání jazyka a v jisté míře i různých žánrových či tematických skupin textů.

Korpus CZESL

První korpus češtiny jako druhého/cizího jazyka vzniká v rámci projektu *Inovace vzdělávání v oboru čeština jako druhý jazyk*.[43] Jsou v něm zařazena data češtiny jako druhého jazyka i jako jazyka cizího.

a) Velikost korpusu a zaměření sběru materiálů

Velikost korpusu byla předběžně stanovena na 2 miliony slov. V porovnání s jinými jazyky – kromě angličtiny – je to rozsah mimořádný.

Sběr materiálu byl zahájen v r. 2009, dokončení korpusu se předpokládá počátkem r. 2012. Sběr není omezen prvním jazykem žáků. Přednostně se zaměřuje na ty jazykové a kulturní skupiny, které mohou být z teoretického i praktického hlediska nejzajímavější: (a) na komunitu s ruštinou a dalšími slovanskými jazyky jako jazyky prvními a (b) s vietnamštinou a dalšími dálněvýchodními jazyky jako prvními jazyky. Ruština a další slovanské jazyky reprezentují jazyky češtině blízce příbuzné; ruština je navíc prvním jazykem významné nové menšiny v ČR. Vietnamština a další dálněvýchodní jazyky reprezentují jazyky nepříbuzné a od češtiny výrazně odlišné; vedle toho je vietnamština druhou významnou novou menšinu v ČR.

Vedle češtiny cizinců se projekt zaměřuje rovněž (c) na češtinu žáků romských. U romských dětí češtinu většinou nelze označit za jazyk druhý. Neplatí to však obecně a navíc je zřejmé, že mezi majoritní společností a romskou komunitou jsou kulturní rozdíly, které mohou osvojování češtiny brzdit a připomínat situaci osvojování jazyka druhého. Romský subkorpus se od ostatních odlišuje i tím, že je zaměřen na školní mládež, je budován vyváženě s ohledem na věk žáků a typ školy a že se ovládání češtiny nehodnotí podle stupnice SERR: jde tedy o akviziční korpus z hlediska L1 a L2 nejednoznačné povahy.

Kromě uvedených jazykových a kulturních komunit obsahuje vstupní materiál i projevy dalších jazykových skupin; jejich prvními jazyky jsou především jazyky západoevropské, zvl. němčina, ale i arabština a další. Půjde tedy o korpus multilingvní, který může být členěn na dílčí subkorpusy podle prvního jazyka mluvčích, podle úrovně ovládání jazyka i podle řady dalších parametrů. Nepředpokládá se vyrovnané zastoupení jednotlivých prvních jazyků.

43 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. Hlavním koordinátorem projektu je K. Šebesta, sběr materiálu u cizinců koordinuje M. Hrdlička a T. Hrdličková, u Romů Z. Bedřichová a K. Šormová, přípravu anotací a vlastní anotování Vl. Petkevič, anotační program připravil J. Hana. Řešení se účastní několik pracovišť obou univerzit i řada studentů doktorského, magisterského i bakalářského studia a spolupracují četná pracoviště neakademická, především školy a občanská sdružení.

Pokud jde o časové zaměření sběru, můžeme CZESL označit za smíšený: bude obsahovat jak longitudinální data získaná od týchž mluvčích v průběhu jednoročního sledování, tak data pseudolongitudinální (tedy data získaná od mluvčích s různou úrovní ovládání jazyka).

Úroveň ovládání jazyka u respondentů (s výjimkou Romů) sahá od začátečníků až po nejvyšší úrovně a rozlišuje se podle systematiky SERR; většina materiálu bude pravděpodobně odpovídat úrovni B2. Romský subkorpus bude odstupňován, jak uvedeno, primárně podle ročníku a podle typu školy.

Od všech sledovaných komunit se sbírají projevy psané i mluvené. Podíl mluvených projevů bude představovat v první fázi přibližně 1/20 korpusu, postupně se však bude zvyšovat, až dosáhne cca 1/4 celého objemu korpusu.

Žánrově i tematicky jde o materiál smíšený, u projevů psaných převládají eseje psané při školních zkouškách, u projevů mluvených elicitovaná interview na obecná témata.

b) Zaznamenávané parametry

K jazykovým datům zařazovaným do korpusu CZESL se zaznamenávají dvě sady sdružených parametrů: (a) parametry pisatele/mluvčího a místa sběru a (b) parametry textu a situace jeho vzniku a záznamu. V celkovém souhrnu je těchto proměnných více než 30; některé z nich se však uplatňují pouze omezeně. Počty sledovaných proměnných a jejich skladbu uvádíme v následujících dvou tabulkách:

Počty sledovaných proměnných v korpusu CZESL

	Cizinci	Romské děti
Písemný materiál	27 proměnných	29 proměnných
Mluvený materiál	31 proměnných	34 proměnných

Sledované proměnné v korpusu CZESL

Parametry vázané k žákovi a místu sběru	Parametry vázané k textu a k situaci jeho vzniku a získán
Pohlaví	Typ textu (rukopis vs. počítač) (PSANÝ)
Věk	Povolen slovník? (PSANÝ)
První jazyk	Je text/projev součást zkoušky?
Studium češtiny – místo (CIZINCI)	Pokud ano, typ zkoušky
Studium češtiny – doba (CIZINCI)	Zadán čas?
Studium češtiny – intenzita (CIZINCI)	Pokud ano, kolik minut?
Studium češtiny – učebnice (CIZINCI)	Zadán rozsah textu? (PSANÝ)
Úroveň češtiny podle SERR (CIZINCI)	Pokud ano, jak velký? (PSANÝ)

Znalost dalších jazyků	Téma (název)
Bilingvnost (CIZINCI)	Typ tématu (obecné vs. speciální)
Délka pobytu v ČR (CIZINCI)	Téma podle zadání (zadané, volné, výběr, jiné)
Umí někdo v rodině česky? (CIZINCI)	Žánr/slohový postup
Pokud ano, kdo? (CIZINCI)	Žánr/slohový postup podle zadání (zadaný vs. volný)
Mluví žák romsky? (ROMOVÉ)	Způsob elicitace a řízení projevu/psaní
Pokud ano, kdy a kde? (ROMOVÉ)	Forma projevu (monolog vs. dialog; účastníci) (MLUVENÝ)
Mluví někdo blízký romsky (ROMOVÉ)	Počet mluvčích (MLUVENÝ)
Pokud ano, kdo? (ROMOVÉ)	Stupeň formálnosti (MLUVENÝ)
Poznámky k žákovi (ROMOVÉ)	Situace při nahrávání (MLUVENÝ)
Typ navštěvované školy (ROMOVÉ)	Míra připravenosti (MLUVENÝ)
Prostředí pořízení materiálu (ROMOVÉ)	První jazyk sběrače[44] (MLUVENÝ)
Místo sběru (ROMOVÉ)	Věk sběrače (MLUVENÝ)
Velikost sídla (ROMOVÉ)	Vztah sběrače k respondentům (MLUVENÝ)
Region (ROMOVÉ)	
Nářeční oblast (ROMOVÉ)	
Sociálně vyloučená lokalita? (ROMOVÉ)	

c) Chybové značkování CZESL

Korpus bude vybaven lingvistickým i chybovým značkováním. Anotační program FEAT (autorem je J. Hana) i postup anotace (zajišťuje V. Petkevič s kolektivem spolupracovníků) byly navrženy s ohledem na specifické rysy češtiny (bohatá morfologie, volný slovosled) a na skutečnost, že anotovány mají být pokud možno i výrazně chybové texty začátečníků a mluvčích s prvním jazykem typologicky vzdáleným.

Anotace je připravena jako anotace vícerovinná, s tím, že jednu rovinu tvoří původní text po přepisu a standardizaci (jde mj. o zajištění anonymity textu). Při jeho převedení na druhou rovinu jsou provedeny takové opravy, aby vznikl řetězec v češtině existujících slov a slovních tvarů, při převodu na třetí rovinu opravy, které vedou ke vzniku gramaticky správných větných celků (tedy např. opravy chyb ve shodě apod.), resp. textů (opravují se i některé chyby, které rámec větného celku přesahují, např. odkazování).

Na obou mezirovinách – mezi rovinami první a druhou, resp. druhou a třetí – jsou zaznamenány chybové i lingvistické značky. Chybová taxonomie kombinuje zřetel k jazykovým rovinám, zvl. k morfosyntaktickým kategoriím, s klasickým formálním popisem chyby (pro-

44 Hodnoty: čeština, stejný jako student, jiný (jaký).

středek chybí, je navíc atd.). Anotace směřuje k systematičnosti, tj. není omezena na některý typ chyb nebo některé jazykové jevy. V budoucnu se počítá s jejím doplněním jemnější anotací speciální, která bude zaměřena úže podle konkrétní výzkumné či pedagogické potřeby.

Postup anotace bude kombinovat emendaci a anotaci manuální s anotací automatickou. Anotační model je vybudován tak, že bude možné uplatnit ho u akvizičních korpusů obecně, ne pouze u korpusů L2.

d) Výzkumné využití a srovnávací korpus

Pro žákovské korpusy (a pro akviziční korpusy obecně) je typické dvojí hlavní výzkumné využití (např. Granger, 1998; a řada dalších autorů).

První navazuje na tradici někdejší chybové analýzy. Korpusy dovolují vyhledávat „chyby" v užívání jazyka (chápané jako odchylky od normy). Na rozdíl od dřívějších sbírek jazykových chyb akviziční korpusy neredukují užívání jazykových jevů na užití chybná, ale dovolují zahrnout do analýzy i užití korektní. Chybová analýza s oporou o korpus podává tedy komplexnější obraz jazykové praxe žáků v situacích s danými parametry.

Druhý způsob využití navazuje na starší tradici kontrastivní lingvistiky. Kontrastivní analýza opřená o korpus se však zaměřuje nikoli na porovnávání výchozího jazyka s jazykem cílovým, nýbrž na porovnávání mezijazyků (proto bývá označována jako kontrastivní analýza mezijazyka – CIA). Může jít o porovnávání mezijazyků různých skupin žáků (např. žáků s různými prvními jazyky, různou úrovní ovládání jazyka apod.), popř. mezijazyka s jazykem výchozím nebo s jazykem cílovým (přesněji řečeno, porovnává se užívání mezijazyka a jazyka výchozího nebo cílového).

Pro účely porovnávání s cílovým jazykem se využívají srovnávací korpusy různého typu. Nemohou to být běžné korpusy synchronní, protože ty se žánrovou skladbou i tematicky od žákovských korpusů výrazně liší. Některé žákovské korpusy (např. ICLE) to řeší vybudováním srovnávacího korpusu rodilých mluvčích. Korpus CZESL s tím rovněž částečně počítá (zejména u subkorpusu romských žáků a subkorpusu mluveného), zčásti se však počítá s využitím korpusu SKRIPT jako korpusu srovnávacího.

Závěr

Akviziční korpusy znamenají nový prvek v korpusové lingvistice a podstatně mění podobu dnešního výzkumu osvojování jazyka. Přinášejí velké objemy autentických a/nebo klinicky elicitovaných jazykových dat zpracovatelných v reálném čase, a umožňují tak zkoumat proces osvojování jazyka s využitím přirozené jazykové produkce žáků v rozsahu, který je jinými cestami nedosažitelný.

Jejich využitelnost roste s vybaveností lingvistickým a chybovým značkováním a s přesným evidováním extralingvistických údajů o situaci vzniku a zpracování korpusových dat. Extralingvistické značkování umožňuje vytvářet dílčí subkorpusy založené na různých kombinacích situačních i osobnostních proměnných a jejich dalekosáhlé porovnávání.

Akviziční korpusy pro češtinu, které se v současné době připravují, se řadí svou velikostí i dalšími parametry k největším neanglickým korpusům svého druhu ve světovém měřítku.

Literatura:

AIJMER, Karin (ed.) (2009): *Corpora and Language Teaching*. Amsterdam – Philadelphia: John Benjamins Publishing Company.

ANDROUTSOPOULOS, Jannis K. – SCHOLZ, Arno (eds.) (1998): *Jugendsprache – langue de jeunes – youth language*. Frankfurt am Main: Peter Lang.

ANDROUTSOPOULOS, Jannis K. – SCHOLZ, Arno (2008): On the recontextualization of hip-hop in European speech communities: a contrastive analysis of rap lyrics. In: Peter Schlobisnki – Niels-Christian Heins (eds.), *Jugendliche und 'ihre' Sprache. Sprachregister, Jugendkulturen und Wertesysteme*. Oldenburg: Westdeutscher Verlag, s. 63–100.

BEHRENS, Heike (2008): Corpora in language acquisition research: History, methods, perspectives. In: Heike Behrens (ed.), *Corpora in Language Acquisition Research*. Amsterdam – Philadelphia: John Benjamins Publishing Company, s. XI–XXX.

BELZ, Julie A. – VYATKINA, Nina (2008): The pedagogical mediation of a developmental learner corpus for classroom-based language instruction [on-line]. Dostupné z WWW: <http://llt.msu.edu/vol12num3/belzvyatkina>.

BERMAN, Ruth A. (ed.) (2004): *Language Development across Childhood and Adolescence*. Amsterdam – Philadelphia: John Benjamins Publishing Company.

ČERMÁK, František – KRÁLÍK, Jan – KUČERA, Karel (1997): Recepce současné češtiny a reprezentativnost korpusu. *Slovo a slovesnost*, 58, s. 117–124.

ČERMÁK, František – SCHMIEDTOVÁ, Věra (2004): Český národní korpus – základní charakteristika a širší souvislosti. *KNIHOVNA. Library Revue*, 15, s. 152–168.

ČMEJRKOVÁ, Světla (2000): Jazyk pro druhé pohlaví. In: František Daneš et al., *Český jazyk na přelomu tisíciletí*. Praha: Academia, s. 146–158.

DÍAZ-NEGRILLO, Ana – FERNÁNDEZ-DOMÍNGUEZ, Jesús (2006): Error Tagging Systems for Learner Corpora. *RESLA*, 19, s. 83–102.

ELLIS, Rod (2008): *Second Language Acquisition*. Oxford: Oxford University Press.

ELLIS, Rod – BARKHUIZEN, Gary (2005): *Analysing Learner Language*. Oxford: Oxford University Press.

GRANGER, Sylviane (ed.) (1998): *Learner English on Computer*. London – New York: Longman.

GRANGER, Sylviane (2009): The contribution of learner corpora to second language acquisition and foreign language teaching. A critical evaluation. In: Karin Aijmer (ed.), *Corpora and Language Teaching*. Amsterdam – Philadelphia: John Benjamins Publishing Company, s. 13–32.

HOFFMANNOVÁ, Jana (1998): Pařani a gamesy. *Naše řeč*, 81, s. 100–111.

INGRAM, David (1994): *First Language Acquisition. Method. Description and Explanation*. Cambridge: Cambridge University Press.

KADLECOVÁ, Kateřina (2006): *Ženské časopisy pro pokročilé. Literární interpretace a kritická analýza diskurzu měsíčníku Marianne*. Liberec: Bor.

KADLECOVÁ, Kateřina (2007): *Dívky a Girls podle českých dívčích časopisů. Jazyk, ideologie, publikum a jeho přístup*. Liberec: Bor.

KALA, Miroslav – BENEŠOVÁ, Marie (1976): *Jak mluví vaše dítě?* Olomouc: Pedagogická fakulta Univerzity Palackého v Olomouci.

KALA, Miroslav – BENEŠOVÁ, Marie (1989): *Písemný a mluvený projev žáků ZŠ*. Olomouc: Pedagogická fakulta Univerzity Palackého v Olomouci.

KOTSINAS, Ulla-Britt (1998): Language contact in Rinkeby, an immigrant suburb. In: Jannis K. Androutsopoulos – Arno Scholz (eds.), *Jugendsprache – langue de jeunes – youth language*. Frankfurt am Main: Peter Lang, s. 125–148.

KOTSINAS, Ulla-Britt (2005): *Invandrarsvenska*. Uppsala: Hallgren & Fallgren Studieförlag.

LÜDELING, Anke et al. (2006): Multi-level error annotation in learner corpora [on-line]. Proceedings of the Corpus Linguistics 2005 Conference. Birmingham, United Kingdom, 14–17 July [cit. 09-10-10]. Dostupné z WWW: <http://www.corpus.bham.ac.uk/PCLC/Falko-CL2006.doc>.

MACWHINNEY, Brian (2000): The CHILDES Project: Tools for Analyzing Talk – Electronic Edition. Part 1: The CHAT Transcription Format. Mahwah, NJ: Lawrence Erlbaum Associates.

MACWHINNEY, Brian (2008): Enriching CHILDES for morphosyntactic analysis. In: Heike Behrens (ed.), *Corpora in Language Acquisition Research*. Amsterdam – Philadelphia: John Benjamins Publishing Company, s. 165–197.

McENERY, Tony – XIAO, Richard – TONO, YuKio (2006): *Corpus-Based Language Studies. An advanced resource book*. London – New York: Routledge.

MALÁ, Michaela (2008): *Jazyk hip hopu a rapu*. Diplomová práce. Praha: FF UK.

NESSELHAUF, Nadja (2005): *Collocations in a Learner Corpus*. Amsterdam – Philadelphia: John Benjamins Publishing Company.

NEWERKLA, Stefan Michael (1999): Jazyk vybraných domovských stránek českých studentů a mladých absolventů vysokých škol v síti 'World Wide Web'. *Naše řeč*, 82, s. 127–138.

NIPPOLD, Marilyn A. (ed.) (1988): *Later Language Development: Ages Nine through Nineteen*. Austin: PRO-ED, Inc.

NIPPOLD, Marilyn A. (1998): *Later Language Development. The School-Age and Adolescent Years*. Austin: PRO-ED, Inc.

PRAVEC, Norma A. (2002): Survey of learner corpora. *ICAME Journal*, 26, s. 81–114.

PŘÍHODA, Václav (1963): *Ontogeneze lidské psychiky I*. Praha: SPN.

ROMAINE, Suzanne (1984): *The Language of Children and Adolescents*. Oxford: Basil Blackwell Publisher.

ROMITI, Sara (1998): The Mural Writings of the Young in Rome. In: Jannis K. Androutsopoulos – Arno Scholz (eds.), *Jugendsprache – langue de jeunes – youth language*. Frankfurt am Main: Peter Lang, s. 281–304.

SANDQVIST, Carin – TELEMAN, Ulf (eds.) (1989): *Språkutveckling under skoltiden*. Lund: Studentlitteratur.

SCHLOBINSKI, Peter – KOHL, Gaby – LUDEWIGT, Irmgard (1993): *Jugendsprache. Fiktion und Wirklichkeit*. München: Wetsdeutscher Verlag.

SINCLAIR, John (1996): Preliminary recommendations on Corpus Typology [on-line]. Dostupné z WWW: <http://www.ilc.pi.it/EAGLES96/corpustyp/corpustyp.html>.

STENSTRÖM, Anna-Brita – ANDERSEN, Gisle – HASUND, Ingrid Kristine (2002): *Trends in Teenage Talk. Corpus compilation, analysis and findings*. Amsterdam – Philadelphia: John Benjamins Publishing Company.

SVOBODA, Karel (1972): *Stylizační výcvik ve spisovné češtině a rozvoj myšlení* 1. Praha: Univerzita Karlova v Praze.

SVOBODA, Karel (1973): *Stylizační výcvik ve spisovné češtině a rozvoj myšlení* 2. Praha: Univerzita Karlova v Praze.

SVOBODOVÁ, Diana – KAVALOVÁ, Eva (1999): O jazyce autorů graffiti. *Naše řeč*, 82, s. 245–254.

ŠEBESTA, Karel (1997): Syntax nedostatků v písemných projevech žáků gymnázií v 1. a 4. ročníku studia. Disertační práce. Praha: FF UK.

ŠULC, Michal (2001): Tematická reprezentativnost korpusů. *Slovo a slovesnost*, 62, s. 53–61.

TAGNIN, Stella E. O. (2010): A multilingual learner corpus in Brazil [on-line]. Dostupné z WWW: <http://www.fflch.usp.br/dlm/comet/artigos/A%20multilingual%20learner%20corpus%20in%20Brazil.pdf>.