Antonia Ordulj
Sveučilište u Zagrebu, Filozofski fakultet, Odsjek za kroatistiku
Gordana Hržica
Sveučilište u Zagrebu, Edukacijsko-rehabilitacijski fakultet, Odsjek za logopediju

Obnavljanje Hrvatskog korpusa dječjega jezika

Updates in Croatian Corpus of Child Language

Stručni rad UDK: 811.163.42'232

SAŽETAK

Hrvatski korpus dječjega jezika (HKDJ - Kovačević, 2002) sadrži zapise spontane interakcije troje djece s odraslim govornicima iz njihove obiteljske sredine. Nastao je longitudinalnim praćenjem u razdoblju dječjega usvajanja jezika u razdoblju od približno godinu dana do tri godine. HKDJ dio je Svjetske baze dječjega jezika CHILDES (http://childes.psy.cmu.edu/data/Slavic/Croatian) te je kodiran u skladu sa sustavom CHAT. Posljednjih je godina postojeći HKDJ podvrgnut detaljnoj analizi kako bi bio u skladu s promjena i novostima iz sustava CHILDES te kako bi postao pristupačniji korisnicima. Cilj je ovoga rada dati pregled postupka kodiranja HKDJ-a u skladu s pravilima sustava CHAT te dobiti uvid u promjene koje se u prvom redu odnose na sustavno kodiranje pogrešaka, a potom i povezivanje transkripata s audiozapisima pomoću *transcriber metode* opisane unutar sustava CHILDES. Sve je ovo omogućilo bolju dostupnost HKDJ-a (sada se može pregledavati i bez instalacije specijaliziranih programa) te dostupnost zvučnih datoteka (mogu se preslušavati izravno iz transkripta i to čitave ili iskaz po iskaz).

Ključne riječi:
CHILDES ▪ HKDJ ▪ CHAT ▪ kodiranje ▪ povezivanje audiozapisa

ABSTRACT

Croatian corpus of child language (CCCL – Kovačević, 2002) consists of language samples of spontaneous interaction of three children and adult speakers in their family environment. The corpus was created by longitudinal monitoring during the period of children's language acquisition from the onset of speech to approximately three years. CCCL is part of the Child Language Data Exchange System CHILDES (http://childes.psy.cmu.edu/data/Slavic/Croatian) and is encrypted in accordance with the CHAT system. In recent years the existing CCCL was subjected to a detailed analysis in order to make it consistent with the changes and developments of the CHILDES system. The aim of this paper is to give an overview of the encoding of CCCL in accordance with the rules of the CHAT system and to gain an insight into the changes that are primarily related to systematic coding of errors. During this process, transcripts have been connected with audio records by using the *transcriber method* described within the CHILDES system. All this allows for the easier access to the CCCL in general (which can now be viewed without installing specialized applications), and access to sound files of each transcript, available both as whole or related to a specific utterance.

Keywords:
CHILDES ▪ HKDJ ▪ CHAT ▪ coding ▪ connecting audio records

Korespondencija: *Antonia Ordulj, Sveučilište u Zagrebu, Filozofski fakultet, Odsjek za kroatistiku, Croaticum – Centar za hrvatski kao drugi i strani jezik, Ivana Lučića 3, 10000 Zagreb, antoniasvetic@gmail.com*

UVOD

Tehnike i metode istraživanja usvajanja prvoga jezika mnogobrojne su i imaju dugu povijest. Jedan od najznačajnijih i najraširenijih pristupa prikupljanja podataka stvaranje je korpusa dječjega govornoga jezika u svakodnevnoj spontanoj uporabi. Korpus je zbirka pisanih ili govorenih tekstova koji predstavljaju određeni oblik ili uporabu nekoga jezika te su namijenjeni lingvističkoj analizi (Crystal, 1998). Riječ je o reprezentativnoj zbirci tekstova u kojima se ogleda jezična raznolikost pojedine jezične zajednice. Zahvaljujući korpusima dječjega jezika moguće je dobiti uvid u razvojne obrasce i stilove usvajanja kod djece.

Povijest uporabe jezičnih korpusa u proučavanju dječjega jezika započinje dnevničkim zapisima tijekom 19. stoljeća, a ovakva se metodologija zadržava sve do 20. stoljeća. Mnogi dnevnici, pored jezičnih informacija, sadrže detaljne informacije o cjelokupnom dječjem razvoju. Tijekom pedesetih i šezdesetih godina prošloga stoljeća, brojni su se istraživači usmjerili na pojedinačne aspekte jezičnoga razvoja pa su nastale korisne dnevničke studije koje nam pružaju informacije o pogreškama i poopćavanjima ili pak uporabi glagola u dječjem jeziku (Behrens, 2008). Pojavom modernih tehnologija došlo je i do promjena na području korpusnih istraživanja dječjega jezika. Audio- i videosnimke jezične interakcije između djeteta, njegovih roditelja i osoba iz obiteljskoga okruženja omogućile su prijenos podataka na računalo, sustavnost te općenito bržu analizu dječjega jezika. Dodatno, omogućeno je snimanje djece različitoga socioekonomskoga podrijetla te uvid u kontekst prikupljenih podataka čime su dnevničke studije bile ograničene jer su gotovo bez iznimke uključivale praćenje djece osoba iz akademske zajednice. No, iako je razvoj računalne tehnologije omogućio brojne pogodnosti, još su uvijek postojale metodološke neujednačenosti u prikupljanju, transkribiranju i kodiranju jezičnih zapisa.

CHILDES

Kako bi se uklonile metodološke prepreke, olakšala analiza spontanoga govora, postigla ujednačenost prikupljenih podataka dječjega jezika te omogućila daljnja istraživanja, nastaje sustav za razmjenu podataka dječjega jezika CHILDES (*Child Language Data Excange System*). Nositelji projekta bili su Brian MacWhinney i Catherine Snow, a realizacija je započela u siječnju 1984. (MacWhinney, 2013; MacWhinney, 2008). Velik broj jezičnih korpusa dječjega jezika danas se nalazi u svjetskoj banci dječjega jezika CHILDES koji ima više od 130 korpusa na više od 20 svjetskih jezika i 1500 članova. Više od 3000 znanstvenih radova objavljeno je analizom podataka iz baze CHILDES (više na: http://talkbank.org/info/usage/childes bib.pdf). Korpusi CHILDES-a javno su dostupni. Korisnik CHILDES-a može postati svatko tko prihvati pravila i uvjete uporabe, a korisnike se potiče i da sudjeluje u daljnjem razvoju i nadogradnji programa i kodova sustava.

Pored podataka djece urednoga jezičnoga razvoja, u CHILDES-u se nalaze i korpusi djece s jezičnim poteškoćama, Downovim sindromom i autizmom, zapisi dvojezične djece i pripovjedni korpus što istraživačima omogućuje sustavno proučavanje i tumačenje dječjega jezika, rijetkih jezičnih pojava, usporedbu podataka i međujezične analize na temelju prijepisa (Kuvač i Palmović, 2007; MacWhinney, 2013). Iako je CHILDES primarno namijenjen istraživačima dječjega jezika, u novije vrijeme sve se više rabi i u dijagnostičke svrhe.

Većina je jezičnih korpusa unutar CHILDES-a primjer spontanoga dječjega jezika snimljenoga u obiteljskoj sredini. Za nastanak takvih korpusa vrlo je važno osigurati prirodnu i ležernu atmosferu kako bi dijete što prije zaboravilo na prisutnost istraživača ili opreme za snimanje. Vrsta podataka ovisi ponajprije o željama i potrebama samoga istraživača, ali i ljudskim, materijalnim i financijskim resursima. Ipak, ključno je prikupiti podatke iz kojih će se moći izvesti opći zaključci i primjerena provjera postavljenih pretpostavaka. Iako ne postoje jasna pravila unutar sustava CHILDES o učestalosti snimanja djeteta, uobičajena je praksa snimati 45 minuta ili sat vremena jednom do dva puta tjedno. Naravno, uvijek postoji mogućnost da se longitudinalno istraživanje nadopuni i nekom drugom metodologijom koja će osigurati još precizniji uvid u neku pojavu ili razdoblje u dječjem jezičnom razvoju (Kuvač i Palmović, 2007; usp. detaljnije o uzimanju uzoraka dječjega jezika i u: Kelić i sur., 2012).

Unutar sustava CHILDES razvijen je sustav kodiranja CHAT te program za računalnu obradu zapisa govornoga jezika CLAN. CHAT (*Codes for the Human Analyses of Transcripts*) služi za oblikovanje prijepisa te se sastoji od kodova kojima se transkribira i kodira zapis dječjega jezika dok računalni program CLAN (*Computerized Language Analysis*) omogućava fonološku, morfološku i sintaktičku analizu zapisa (MacWhinney, 2013). Pored navedenih programa i različitih alata kojima se neprestano usavršava pretraživanje i mogućnost obrade korpusa, s razvojem računalnih tehnologija omogućeno je i povezivanje originalnoga audio ili videozapisa s transkriptom. Na taj način istraživačima je omogućen izravan uvid u interakciju između djeteta i njegovih bližnjih, a transkript je podložan neprestanoj provjeri i analizi mogućih nepravilnosti (Behrens, 2008). Unutar sustava CHILDES razvijeno je pet metoda kojima je moguće povezivanje originalnoga audio ili videozapisa s transkriptom (*sonic mode, transcriber mode, video mode, sound walker, time mark editing*), a odabir ovisi o svrsi svakoga pojedinačnoga istraživanja i željama istraživača (MacWhinney, 2013).

Svi podaci iz korpusa autorsko su djelo istraživača pa je potrebno navoditi točne podatke prilikom citiranja ili korištenja podataka. Svi korpusi i potrebni podaci nalaze se na internetskoj stranici http://childes.psy.cmu.edu/ , a za što lakše snalaženje unutar CHILDES-a, na istoj je adresi moguće pronaći priručnike s uputama za uporabu sustava i transkripciju.

Hrvatski korpus dječjega jezika

Sustavan rad na Hrvatskom korpusu dječjega jezika (HKDJ – Kovačević, 2002) započinje tijekom 1999. godine osnutkom Laboratorija za psiholingvistička istraživanja POLIN Sveučilišta u Zagrebu. Riječ je o korpusu koji je nastao metodom kontinuiranoga uzorkovanja u razdoblju dječjega usvajanja jezika. HKDJ nastaje snimanjem djece do tri puta mjesečno po četrdeset i pet minuta prilikom spontane komunikacije i interakcije djece s ukućanima u obiteljskoj sredini. Iako je snimano više djece, HKDJ obuhvaća zapise samo njih troje - Antonije, Marine i Vjerana

Navedeni su zapisi u potpunosti transkribirani te su dostupni kao dio baze CHILDES na internetskoj adresi http://childes.psy.cmu.edu/data/Slavic/Croatian. Ispitanici koji su sudjelovali u longitudinalnom praćenju tijekom snimanja materijala za HKDJ, govornici su koji tek usvajaju jezik. Riječ je o troje djece, Antoniji, Marini i Vjeranu, koji su u neprestanoj interakciji s odraslim govornicima iz svoje obiteljske sredine.

Dvije se skupine govornika, osim prema kronološkoj dobi, razlikuju i idiomom kojim se služe. Tijekom ovoga istraživanja naglasak je bio na idiomu *dječjega jezika* budući da ukazuje na razvojne obrasce i neprestane promjene prilikom usvajanja hrvatskoga jezika. Longitudinalna istraživanja najčešće predstavljaju napor kako za istraživača tako i za same ispitanike i njihovu okolinu (osobito roditelje djece koja se snimaju) pa je pored jasnih istraživačkih ciljeva, potrebno uzeti u obzir i niz drugih čimbenika (socijalni status, obrazovanje roditelja, početak usvajanja jezika, različite idiome i narječja) koji mogu utjecati na buduću analizu podataka.

CILJ RADA

Neprestani razvoj i usavršavanje programa i kodova sustava CHILDES jedna je od temeljnih zadaća svakoga korisnika. Tijekom akademske godine 2011./2012. tim je stručnjaka iz Laboratorija za psiholingvistička istraživanja Sveučilišta u Zagrebu proveo detaljnu analizu postojećega HKDJ-a s ciljem da se hrvatski korpus dodatno modernizira i bude u tijeku s promjenama unutar sustava CHILDES. Promjene su se unutar HKDJ-a odnosile ponajprije na ujednačavanje i ispravljanje postojećih kodova, ususstavljivanje kodova za pogreške i povezivanje originalnoga audiozapisa s transkriptima. Stoga je cilj ovoga rada dati pregled glavnih obilježja i kodova sustava CHAT na primjeru hrvatskoga korpusa na CHILDES-u i promjena kojima je HKDJ bio podvrgnut. Promjene su bile dvojake. Prvo, ususstavljeni su rabljeni kodovi za označavanje dijelova teksta, s posebnim naglaskom na sustavno kodiranje pogrešaka. Drugo, takvi su uređeni transkripti povezani s audiozapisima.

Promjene u kodiranju zapisa govornoga jezika

Izgled zapisa govornoga jezika

CHAT je standardizirani sustav za transkripciju CHILDES banke dječjega jezika te su, shodno nazivu, sve datoteke s prijepisom zvučnih zapisa označene odgovarajućom oznakom *.cha*. Svaka CHAT datoteka sadrži dva glavna dijela: zaglavlje i prijepis. Da bi program u potpunosti ispravno radio, u zaglavlju je potrebno navesti osnovne podatke o zapisu: jezik, sudionike, djetetovu dob i spol, datum, naziv datoteke i opis situacije (primjer 1).

Prijepis čine iskazi svakoga pojedinoga sudionika. Vrlo je važno da svaki iskaz bude u zasebnom retku, a da bi se osigurala čitljivost i organiziranost prijepisa, razlikuju se glavni ili sadržajni redak i zavisni redak. Glavni ili sadržajni redak prijepisa započinje zvjezdicom, troslovnom oznakom govornika i dvotočkom, a na kraju retka dolazi interpunkcijski znak (*ANT:). Zavisni redak prijepisa započinje oznakom **'%'**, a nadopisuje ga prepisivač kako bi dao komentar ili opisao situaciju (*%com* ili *%sit*), naveo kome se govornik obraća (*%add*), označio fonološke ili morfološke kodove (*%pho ili %mor*) ili kodirao pogrešku govornika (*%err*) (primjer 2).

Unutar prijepisa moguće su situacije kada se zavisni redak odnosi na više sadržajnih redaka. Tada započinje oznakom @ i odgovarajućim kodom, npr. @*Comment* – ako se želi komentirati veći broj iskaza, @*New Episode* – ako kreće novo snimanje nakon kraćega prekida, @*Bg* označava početak, a @*Eg* kraj onih situacija koje isključujemo iz analize (npr. pjevanje ili pričanje priče) ako je cilj analizirati samo spontani govor djeteta (primjer 3 i 4).

Primjer 1. Zaglavlje datoteke CHAT iz Hrvatskoga korpusa

@Begin
@Language: hr
@Participants: ANT Antonija Target_Child, DRA Draženka Mother
@Sex of ANT: female
@Birth of ANT: 31-JUL-1993
@Age of ANT: 1;3.3
@Date: 3-NOV-1994
@Filename: antbla01.cha
@Coding: CHAT
@Situation: Antonija se u sobi igra s lutkicom i priča

Primjer 2. Prikaz glavnoga i zavisnoga retka s opisom situacije

*DRA: mamici hoćeš ići?
*DRA: evo ga.
*DRA: hajde tu.
%com: uzima je u krilo

Primjer 3. Prikaz novoga snimanja nakon kraćega prekida

@New Episode
*DRA: što je to?
*ANT: va [/] va [/] va
%com: pokazuje magnetofon

Primjer 4. Prikaz zavisnoga retka s većim brojem sadržajnih redaka

@Bg: pjevanje
*VJE: ti si ljubav.
*MEL: jedina.
*VJE: jedino jedina.
@Eg: pjevanje

Ujednačavanje zapisa govornoga jezika.

U HKDJ-u su se opisi situacija nesustavno bilježili u dva zavisna retka, %com i %sit. U obnovi korpusa u retku *%com* upisani su samo komentari vezani uz sam iskaz (kao u

primjeru 2 i 3), a u retku *%sit* upisani su komentari koji se odnose na informacije koje nisu nužne za razumijevanje iskaza za koji su vezane. Na primjer, tako se označava ulazak nove osobe u prostoriju, buka koja onemogućuje komunikaciju i slično. Na taj se način lako može pratiti samo one informacije o kontekstu razgovora koje korisnika korpusa zanimaju. Također, analiza situacijskih informacija postaje lakša jer korisnik na jednostavan način može izdvojiti samo retke čiji ga sadržaj zanima. Popravljene su i neujednačenosti u zaglavljima datoteke.

Kodovi za označavanje riječi u HKDJ-u

Izvanjezični i dodatni jezični elementi te kontekst snimanja opisuju se sustavom kodova koji se nalaze u retku u kojem je iskaz ili u zasebnim recima ispod iskaza na koji se odnose. MacWhinney (2012) je kodiranje definirao kao proces prepoznavanja, analiziranja i bilježenja fenomena transkribiranoga govora. Upotreba kodova ovisi o interesima istraživača što znači da će rijetko koji korpus sadržavati sve kodove. U CHILDES priručniku nalazi se više od stotinu različitih kodova, a najčešći koji su rabljeni u kodiranju HKDJ-a bit će prikazani u ovom poglavlju.

S obzirom na to da su i dječji jezik i govor usmjeren djetetu govorni idiomi, često je potrebno naznačiti posebnosti leksičkoga materijala koji se rabi, bilo da je riječ o dijalektalizmima ili leksičkim tvorbama karakterističnima za takvu komunikacijsku situaciju (primjerice, brbljanje ili izmišljena dječja riječ). Obilježavanje riječi kodovima (primjer 5) omogućuje da se u daljnjoj analizi takva riječ izuzme ili na poseban način tretira. Primjerice, pri izračunavanju prosječne duljine iskaza u riječima, jedne od mjera jezičnoga razvoja (na primjer, Kuvač i Palmović, 2007), nužno je izuzeti brbljanje jer podaci u protivnom ne bi odgovarali slici jezičnoga razvoja. Ako su riječi sustavno kodirane, pri jednostavnom i automatiziranom izračunu mjera jezičnoga razvoja sve se riječi koje imaju određen kod (u ovom primjeru brbljanje koje se označava kodom @b) mogu iz izračuna izuzeti.

Primjer 5. Označivači riječi u Hrvatskom korpusu

1. Brbljanje:
*DRA: molim maco?
*ANT: mama ma@b ba@b mama

2. Izmišljena dječja riječ i riječ koju je usvojila obitelj:
*MAR: aja@c beba aja@c
*SAN: ko [: tko] aja@f?

3. Izmišljena dječja riječ:
*MAR: ade@b nane@c [= misli na banane]

4. Dijalektalni oblici:
*ANA: ali ne smeš@d [: smiješ] kavu dirati.
*ANT: neću.

5. Onomatopeja:
*MAR: didi@o [= onomatopeja od voziti auto]

6. Pojedinačno slovo:
*RAD: je l(i) vidite kako on r@l (.) r@l izgovara sočno.

7. Imitacija:
*SAN: šta čita?
*MAR: čita@im

8. Slogovanje:
*ANT:da.
*ANT: ta@sly tata.

Pored ovih oznaka, često se rabe i one kojima želimo označiti nerazumljivu (*xxx, xx, www, 0*), ispuštenu (*0word*) ili nedovršenu riječ (). U primjeru 6 može se vidjeti da je dio nedovršene riječi '*čekaj*' stavljen u zagradu, a riječi koje su prepisivaču bile nerazumljive označene su oznakom '*xxx*'.

S obzirom na velik broj nedovršenih riječi, u ponovljenoj se analizi HKDJ-a odlučilo za dosljedno bilježenje nedovršenih riječi u uglatim zagradama, a među najfrekventnije riječi pripadaju: *di* [: gdje], *došo* [: došao], *kolko* [: koliko], *ajde* [: hajde], *oćeš* [: hoćeš], *ko* [: kao], *jel* [: je li], *ak* [: ako], a isti se način bilježenja primijenio i na riječi *ovak, tak* i *onak*. Ovakav način kodiranja zapravo je svakoj od navedenih riječi pridružio jedinstveni oblik prema kojem će ga alati CLAN-a tretirati. Tako će, primjerice, sustav i *di [: gdje]*, i *gdje* i *dje [: gdje]*, prepoznati kao tri pojavnice oblika *gdje*. To omogućuje znatno veću preciznost pri automatiziranoj izradi čestotnih rječnika različnica te pri izračunu mjera leksičkog razvoja (na primjer, omjer različnica i pojavnica (ORP – Kuvač, Palmović, 2007), prema type-token ratio (TTR).

Među ispuštenim riječima najčešće se nalaze oblici pomoćnih glagola, zamjenice ili prijedlozi što se označava oznakom *nula '0'* i slovnom oznakom vrste riječi koja je ispuštena, npr. *0v* (ispušten glagol), *0aux* (ispušten pomoćni glagol kao u primjeru 7), *0ref* (ispuštena povratna zamjenica) ili *0prep* (ispušten prijedlog).

Primjer 6. Označavanje nedovršene '()' i nerazumljive riječi (xxx)

*ANT: ček(aj) [/] ček(aj) (.) ide xxx.

Primjer 7. Ispuštena riječ (0aux)

*VJE: a di [: gdje] 0aux voka@c [/] a di [: gdje] 0aux voka@c?

Ujednačavanje kodova u zapisima.

U ponovnoj se analizi HKDJ-a osobita pozornost posvetila ujednačavanju kodova riječi koje predstavljaju specifične oblike dječjega jezika. To se ponajprije odnosi na dosljedno bilježenje dijalektalnih oblika s obzirom na visoku zastupljenost štokavskoga narječja u interakciji između odraslih govornika i snimane djece te na označavanje brbljanja. Velike su promjene uvedene u označavanje ispuštenih riječi koje često nisu bile uopće označavane. Do neke mjere bila su označena samo ispuštanja pomoćnih glagola. Ovom su obnovom HKDJ-a uvedene i kategorije*0v*, *0n*, *0ref* te *0prep*. Dosljedno označavanje ovakvih ispuštanja omogućuje da se jednostavnom pretragom ili primjenom

naredbe KWAL lako i brzo izdvoje i prouče iskazi u kojima je došlo do ispuštanja, bilo u određenim transkriptima, bilo u potkorpusu jednog djeteta ili u čitavom HKDJ-u. Ovi su podaci važni za proučavanje jezičnoga razvoja jer je ispuštanje funkcionalnih riječi jedna od karakteristika dječjega jezika u njegovom razvojnom procesu, a na ovaj je način vidljivo kada se javlja te kada završava ovo razvojno razdoblje.

Kodovi koji se odnose na iskaz

U Hrvatskom se korpusu koriste i oni kodovi koji se odnose na cijeli iskaz. Tako je moguće označiti iskaz koji nije dovršen (+...), kojega je govornik sam prekinuo (+//.) ili ga je prekinuo sugovornik (+/.) (primjer 8 i 9).

Primjer 8. Govornik sam prekida svoj iskaz

*DRA: ali ipak brojiti zna.
*DRA: tu su +//.

Primjer 9. Govornika je prekinuo sugovornik

*MEL: ma baci mi loptu baci +/.
*VJE: radi mama radi.

Među vrlo čestim kodovima, javljaju se oznake za preklapanje govornikova i sugovornikova iskaza (*[>][<]*), ponavljanje riječi (*[//]*) ili ponavljanje s ispravljanjem (*[///]*), a ponekad i potpuno preoblikovanje iskaza (*[////]*) (primjer 10, 11 i 12). Ako se u iskazu pojavljuje upravni govor, najčešće tijekom prepričavanja priče, označava se navodnim znacima u uglatoj zagradi (*["]*) (primjer 13). Među vrlo čestim kodovima je i oznaka za kraću ili dužu pauzu (primjer 14 i 15).

Primjer 10. Preklapanje govornikova i sugovornikova iskaza

*VJE: vuci [: uči] djecu <mama> **[>]**.
*MEL: <uči djecu> **[<]**.

Primjer 11. Ponavljanje s ispravljanjem

*BLA: i tak [: tako] veliš mama vozi mad **[//]** mazdu.

Primjer 12. Ponavljanje

*VJE: svaki kat ima **[/]** ima stepenice.

Primjer 13. Upravni govor

*DRA:<upomoć [/] upomoć> **["]**
*DRA:<neka mi netko pomogne>**["]**
*DRA:<moj traktor se zaglavio, vikao je hans prestrašeno, a životinje su pomislile, to mu je kazna zato što je sa svojim bučnim traktorom poremetio naš mir.> **["]**

Primjer 14. Kraća pauza

*DRA: tko je to bacio lopticu?
*DRA: bravo (.) evo gol.

Primjer 15. Duža pauza

*ANT: čekaj **[/]** čekaj **(..)** tupu čekaj.

*Ujednačavanje kodova u zapisima***.**

U obnovi HKDJ-a ispravljeni su pogrešno rabljeni kodovi za ponavljanje s ispravljanjem i potpuno preoblikovanje iskaza te umetnuti kodovi za upravni govor. Također, s obzirom na to da je u postojećem HKDJ, pauza bila dosljedno označena prvotnim znakovima '#' za kraću, odnosno '##' za dužu pauzu koji pripadaju starijim inačicama sustava za označavanje CHAT, u ponovljenoj su se analizi stari kodovi za pauzu dosljedno zamijenili novima i to (.) za kraću i (..) za dužu pauzu.

Kodiranje pogrešaka

Kodiranje pogrešaka u Hrvatskom je korpusu moguće na dva načina: u glavnom retku tako da se ispravan oblik stavi u uglatu zagradu iza dvotočke (primjer 16) ili u posebnom retku *'%err – error line'* u kojem se pomoću zasebnih kodova može označiti i vrstu pogreške što prvim načinom nije moguće (primjer 17). U HKDJ-u su pogreške dosljedno bilježene u posebnom *'%err'* retku što omogućuje veću preciznost pri pretraživanju te odmah dostupan opis pogreške. Prvi je način kodiranja uglatom zagradom zadržan samo za izgovorne inačice pojedinih riječi. Među najčešće rabljenim kodovima za pogreške javljaju se kodovi za fonološke, leksičke, sintaktičke i morfološke pogreške te situacije kada je nejasno kako klasificirati pogrešku.

Tablica 1. Kodovi za pogreške u HKDJ-u

$PHO – fonološka pogreška	$CC – pogrešan konsonantski slijed
$LEX – leksička pogreška	
$SYN - sintaktička pogreška	$POS – pogrešan red riječi
$MOR – morfološka pogreška	$PRE – pogrešan prefiks $SUF – pogrešan sufiks $NFL – pogreška u deklinaciji ili konjugaciji $DER – pogreška u tvorbi riječi $REG - poopćavanje
$UNC – nejasno kako klasificirati pogrešku	

Primjer 16. Označavanje pogreške stavljanjem u uglatu zagradu u glavnom retku

*DRA: a **di [: gdje]** ti je seka?
*ANT: a tamo je, tamo u **košajici [: košarici].**

Primjer 17. Označavanje pogrešaka u *error* retku

a) Morfološka pogreška – pogrešan prefiks
*ANT: a ja ću ti podavati [*]
%err: podavati = dodavati $MOR $PRE

b) Morfološka pogreška – pogreška u deklinaciji
*ANT: i ove **[*]** dva.
%err: ove = ova $MOR $NFL

c) Morfološka pogreška – poopćavanje
*ANT: a oćem [*].
%err: oćem = hoću $MOR $REG

d) Fonološka i morfološka pogreška
*ANT: <na fesetu> [*]
%err: na fesetu = u sesvetama $PHO $CC $MOR $NFL

e) Leksička pogreška
*MAR: maže **[*]** (.) mama njegova.
%err: maže = mazi $LEX

f) Sintaktička pogreška – pogrešan red riječi
*MIR: morat ćemo ti malo glazbe naći <sad mi si> **[*]** se jako uozbiljio
%err: sad mi si = sad si mi $SYN $POS

S obzirom na to da je hrvatski jezik morfološki vrlo razvijen te se u morfološki bogatim jezicima infleksijski oblici javljaju vrlo rano, kodiranje pogrešaka u HKDJ-u zahtijevalo je detaljnu i preciznu analizu kako bi se ispravnost označavanja postojećih pogreška dodatno ispitala, a brojne neoznačene pogreške sustavno označile odgovarajućim kodovima.

Uvedene su inovacije u označavanju potrebne za razlikovanje pojava karakterističnih za usvajanje hrvatskoga. Tako su na različit način tretirana preopćavanja fleksijskog morfema (kao u primjeru 'možem') i preopćavanja glagolske osnove (na primjer 'pisam' umjesto 'pišem'). U HKDJ-u su najzastupljenije morfološke pogreške u deklinaciji i konjugaciji te preopćavanja.

Sustavnim kodiranjem moguće je rabeći alat KWAL izdvojiti samo pogreške koje korisnika korpusa zanimaju ili izdvojiti sve pogreške u samo nekim transkriptima, u jednom od podkorpusa ili u čitavom HKDJ-u. Tako je moguće razvojno pratiti, primjerice, obrasce preopćavanja.

Nakon dugotrajnoga procesa transkribiranja i kodiranja na zapis se primjenjuje program CHECK iz programskoga paketa CLAN kako bi se utvrdilo je li zapis u skladu s pravilima sustava CHAT te jesu li svi kodovi i komentari točno upotrijebljeni. Računalo će ispisati svaki redak s pogreškom, a tek kada se one uklone, zapisi su spremni za daljnju obradu i analizu.

Povezivanje zvuka i teksta

Unutar sustava CHILDES moguće je povezivanje tekstualnoga zapisa s originalnim audio ili videozapisom. Tako je transkript podložan neprestanoj provjeri i analizi potencijalnih nepravilnosti jer ga neprestano preslušavaju novi korisnici. Nadalje, zvučni zapis omogućuje lakšu provjeru nerazumljivih riječi, specifičnih oblika dječjega jezika poput brbljanja ili onomatopeja, pogrešaka i slično (Behrens, 2008).

Povezivanje originalnoga audio- ili videozapisa s transkriptom moguće je odabirom jedne od pet metoda (*sonic mode, transcriber mode, video mode, sound walker, time mark editing*). Za povezivanje zvučnih datoteka u HKDJ-u odabrana je *transcriber metoda* s obzirom na to da je puno brža od primjerice *sonic* ili *video metode* te se može koristiti i za povezivanje videozapisa s transkriptom.

Nadalje, *transcriber metoda* pogodna je i za povezivanje zvučne datoteke s postojećim .cha datotekama, ali i za slučajeve kada se iz zvučnih datoteka žele stvoriti novi transkripti (MacWhinney, 2013). S obzirom na to da je HKDJ već transkribiran u skladu s pravilima sustava CHAT, za povezivanje zvučnih datoteka s postojećim tekstualnima bilo je dovoljno slijediti nekoliko jednostavnih koraka opisanih u CHILDES priručniku (MacWhinney, 2013:29).

ZAKLJUČAK

Stvaranje CHILDES korpusa kao reprezentativnoga uzorka omogućilo je istraživačima sustavno praćenje dječjega jezičnoga razvoja i međujezične analize. Druga je prednost što su CHILDES korpusi i programi javno dostupni u elektronskom obliku, neprestano se razvijaju i pružaju brojne mogućnosti ručne i automatske obrade jezičnoga sadržaja.

No, s druge strane stvaranje i održavanje korpusa iznimno je dugotrajan i iscrpljujuć posao, a složenost u transkribiranju i kodiranju jezičnoga materijala zahtjeva velike napore istraživača i sustavno korištenje kodova i simbola kako bi se osigurala jasnoća i čitljivost prijepisa.

Korpus dječjeg jezika kao što je HKDJ prikladan je za dohvat podataka o jezičnom razvoju i međujezične usporedbe, ali i ima potencijale i za uporabu u logopedskoj praksi. Naime, u mnogim je zemljama uzimanje uzorka dječjega jezika uobičajen dijagnostički postupak. Uzorak se transkribira i izračunavaju se mjere jezičnoga razvoja (na primjer, omjer obličnica i pojavnica ili prosječna duljina iskaza – više u Kelić i sur., 2012). Za neke su jezike te mjere standardizirane, no ne i za hrvatski. Ipak, mogu se uspoređivati s onima izmjerenima na korpusima dječjih jezika, ali i s drugim ispitanicima. Posebno su u tu svrhu korisni uzorci koji se uzimaju situacijski – na primjer uporabom slikovnoga predloška ili konvencionalnih pitanja.

Dakle, uzorci govornoga jezika mogu biti raznovrsni, ali sustav za transkripciju (CHAT) i program za obradu transkripata (CLAN) omogućuju znatno jednostavniju analizu transkripata i automatizirano označavanje dijela mjera jezičnoga razvoja.

LITERATURA

1) Bahrens, H. (2008). Corpora in language acquisition research: History, methods, perspectives. U: H. Bahrens (ur.) *Corpora in Language Acquisition Research.* Amsterdam-Philadelphia: John Benjamins Publishing Company, 11-30.
2) Crystal, D. (1998). *The Cambridge Encyclopedia od Language*, Cambridge University Press.
3) Helimann, J. (2010). Myths and Realities of Language Sample Analysis. *SIG 1 Perspectives on Language Learning and Education, 17*, 4-8. doi:10.1044/lle17.1.4.
4) Hržica, G. (2011). *Glagolske kategorije aspekta, vremena i akcionalnosti u usvajanju hrvatskog jezika*. Doktorska disertacija. Zagreb. Filozofski fakultet.
5) Kelić, M., Hržica, G., Kuvač Kraljević, J. (2012). Mjere jezičnog razvoja kao pokazatelji posebnih jezičnih teškoća. *Hrvatska revija za rehabilitacijska istraživanja, 12 (2),* 23-40.
6) Kovačević, M. (2002). *Croatian corpus, CHILDES*. http://childes.psy.cmu.edu/data/Slavic, (4.2.2015.)
7) Kuvač, J., Palmović, M. (2007). *Metodologija istraživanja dječjeg jezika.* Zagreb: Naklada Slap.
8) MacWhinney, B. (2013). *The CHILDES Project: Tools for Analyzing Talk – Electronic Edition. Part 2: The CLAN Programs*. Carnegie Mellon University.
http://childes.psy.cmu.edu (15.12.2014.)
9) MacWhinney, B. (2012). *The CHILDES Project: Tools for Analyzing Talk – Electronic Edition, Part 1: The CHAT Transcription Format.* Carnegie Mellon University.
http://childes.psy.cmu.edu (15.12.2014.).
10) MacWhinney, B. (2008). Enriching CHILDES for morfosyntactic analysis. U: H. Bahrens (ur.) *Corpora in Language Acquisition Research.* Amsterdam-Philadelphia: John Benjamins Publishing Company, 165-197.