UDK 811.163.42’322
Izvorni znanstveni rad
Rukopis primljen 13. IX. 2019.
Prihvaćen za tisak 23. I. 2020.
doi.org/10.31724/rihjj.46.2.24

Nives Mikelić Preradović
Filozofski fakultet Sveučilišta u Zagrebu
Ulica Ivana Lučića 3, HR-10000 Zagreb
nives.mikelic@gmail.com

OZNAČAVANJE POGREŠAKA U CROLTEC-U (RAČUNALNOM UČENIČKOM KORPUSU HRVATSKOG KAO INOG JEZIKA)

U radu je opisana shema za označavanje pogrešaka u CroLTeC-u – prvom računalnom učeničkom korpusu hrvatskog kao inog jezika. Shema označavanja pogrešaka djelomično se temelji na shemi upotrijebljenoj u razvojnom korpusu slovenskog jezika – *Šolaru* i shemi koja se upotrebljava u korpusu *Cambridge Learner* te je prilagođena hrvatskom jeziku. Shema je razvijena kako bi se označio dio tekstova u korpusu *CroLTeC* te kako bi se omogućilo istraživačima i stručnjacima koji se bave proučavanjem hrvatskog kao inog jezika da uoče koji aspekti jezika određenim skupinama učenika uzrokuju najviše poteškoća u usvajanju hrvatskog te da prilagode nastavne materijale različitim skupinama učenika (ne samo s obzirom na njihovo poznavanje hrvatskog jezika nego i s obzirom na materinski jezik).

1. Uvod

Učenički korpusi, tj. elektroničke zbirke tekstova neizvornih govornika određenog jezika, bogat su izvor informacija o jeziku tih govornika.[1] Tekstovi u njima mogu biti označeni metapodatcima, morfosintaktičkim kategorijama, može im se računalno označiti sintaktička struktura, a mogu imati označene i pogreške (Granger 2008: 259–274).

[1] Učeničkim korpusima najčešće se smatraju zbirke tekstova nastale na tečajima stranih jezika i obično sadržavaju tekstove punoljetnih osoba koje uče neki jezik kao drugi ili strani (dakle, učenici nisu osnovnoškolci ili srednjoškolci). Korpusi koji prate jezični razvoj osnovnoškolaca i srednjoškolaca nazivaju se razvojnim korpusima, a radi se o elektroničkim zbirkama tekstova izvornih maloljetnih govornika.

Obrada učeničkog korpusa sastoji se od automatskih i ručnih zadataka. Koraci automatske obrade obično su tokenizacija (rastavljanje teksta na rečenice i riječi), lematizacija (svođenje riječi u tekstu na njihov kanonski oblik) i morfosintaktičko označavanje te sintaktička raščlamba. Nažalost, računalni alati koji izvršavaju te zadatke skloniji su pogreškama pri procesu izrade korpusa nematerinskoga jezika. Koraci ručne obrade mogu uključivati normalizaciju (tj. ispravljanje pogrešaka) i označavanje pogrešaka (tj. klasifikaciju pogrešaka s pomoću određene taksonomije).

U ovom radu predstavljamo sustav označavanja pogrešaka razvijen za hrvatski učenički korpus CroLTeC (engl. ***Cro****atian* ***L****earner* ***Te****xt* ***C****orpus*) koji se sastoji od otprilike milijun pojavnica prikupljenih iz tekstova neizvornih govornika hrvatskog jezika koji su pohađali tečajeve hrvatskog kao inog jezika na svim jezičnim razinama Zajedničkog europskog referentnog okvira za jezike, od A1 (pripremni korisnik) do C2 (vrsni korisnik).

Pokazalo se da je označavanje pogrešaka važan aspekt istraživanja učeničkog korpusa jer pomaže identificirati problematična područja u procesu usvajanja inog jezika (Granger 2003: 465–480), no ključno je upravo za područje računalne obrade prirodnog jezika koje se naziva računalno potpomognuta analiza pogrešaka (Díaz-Negrillo i dr. 2010: 139–154).

U ovom radu opisat će se taksonomija za označavanje pogrešaka u učeničkom korpusu hrvatskog kao inog jezika, kao i prvi rezultati označavanja pogrešaka u korpusu CroLTeC, gdje pogreške nisu samo ispravljene nego su i klasificirane prema toj taksonomiji. Budući da je CroLTeC prvi korpus s označenim pogreškama za hrvatski kao ini jezik, otvaraju se nove mogućnosti u proučavanju hrvatskog kao inog jezika.

2. Učenički korpusi i označavanje pogrešaka

Svrha je učeničkih korpusa omogućiti dubinsku analizu jezika koji učenici usvajaju kao drugi i strani, usporediti jezik neizvornih i izvornih govornika i odstupanje od standarda. Korpusi bi trebali omogućiti uočavanje jezičnih obrazaca, kontrastivnu međujezičnu analizu i računalno podržanu analizu pogrešaka. Uče-

nički korpusi mogu se uspoređivati s korpusima izvornog jezika, što pomaže u praćenju raznih odstupanja koja proizvode neizvorni govornici.

Iscrpan popis više od stotinu trenutačno dostupnih učeničkih korpusa razvijen je u Centru za englesku korpusnu lingvistiku na Université catholique de Louvain.[2] Što se tiče slavenskih jezika, koji imaju bogatu i složenu morfologiju, dostupna su samo tri učenička korpusa: ruski učenički korpus RLC (Rakhilina i dr. 2016: 66–75), slovenski korpus PiKUST (Stritar 2009: 135–152) i češki korpus CzeSL (Hana i dr. 2010: 11–19).

U kontekstu označavanja učeničkih korpusa termin *pogreška* (engl. *error*) općeprihvaćen je u literaturi (Rastelli 2009)[3], a najpoznatiji učenički korpusi engleskog (*International Corpus of Learner English* ICLE), švedskog (*Swedish Learner Language Corpus* SweLL), norveškog (*Norsk andrespråkskorpus* ASK) i portugalskog (*Corpus de Português Língua Estrangeira/Língua Segunda* COPLE2) korpusi su s označenim pogreškama (engl. *error-tagged corpora*). Također, termin *označavanje pogrešaka* (engl. *error annotation*) upotrebljava se još od 1998. (Dagneaux i dr. 1998) za definiranje prvog koraka u računalno potpomognutoj analizi pogrešaka (engl. *computer-aided error analysis – CEA*) na temelju učeničkih korpusa. Cilj je označavanja pogrešaka izgraditi korpus koji može poslužiti kao repozitorij autentičnih podataka o međujeziku učenika. Prema Selinkeru (Selinker 1972), međujezičnost se mijenja s učenikovim napretkom kroz sukcesivne faze stjecanja jezičnih kompetencija i može se promatrati kao individualni i dinamički kontinuum između materinskog i ciljnog jezika učenika.

Što se tiče načina označavanja pogrešaka u učeničkim korpusima, ručno označavanje još je uvijek najčešći postupak i podrazumijeva značajan ljudski napor, iako postoje eksperimenti s automatskim označavanjem pogrešaka gdje računalni sustav ili pomaže anotatorima ili samostalno označava tekstove bez uključenosti ljudskih anotatora (Rosén 2017: 163–180; Rosén, Wirén i Volodina 2018: 181–184; Flor i Futagi 2012).

Pregled najsuvremenijih tehnika ručnog označavanja pogrešaka u učeničkim korpusima (del Río i Mendes 2018: 225–239) pokazuje da još uvijek ne postoji

[2] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

[3] Stemle i dr. (2019) navode da neki autori pogrešku nazivaju *odstupanjem od norme* ili *međujezičnim fenomenom* (Díaz-Negrillo i dr. 2010), neki *neočekivanom uporabom* (Gaillat i dr. 2014), a neki *oblikom koji ne odgovara normi* (Dobric 2015).

standardizirana taksonomija pogrešaka, ali i da je većina sustava označavanja pogrešaka međusobno slična s teorijskog stajališta. Naime, takvi su sustavi namijenjeni pisanim tekstovima i upotrebljavaju otprilike iste jezične razine: pravopis, gramatiku i leksik.

Prvi korak u označavanju pogrešaka u učeničkom korpusu naziva se normalizacijom ili označavanjem ciljne hipoteze[4] (Lüdeling 2015: 142–145), kad se pogrešan izvorni tekst ne zamjenjuje normaliziranom (ispravljenom) verzijom teksta, nego se normalizirani tekst dodaje kao novi sloj uz izvorni tekst (Stemle i dr. 2019). Nakon toga slijedi drugi korak, dodjeljivanje jedne ili više oznaka pogrešci (npr. riječ može biti pravopisno i gramatički neispravna).

U mnogim učeničkim korpusima prvi se korak (normalizacija) i dodjeljivanje oznake pogrešci (drugi korak) kombiniraju u jedan korak, ali postoje korpusi u kojima je tekst samo normaliziran i nisu označene pogreške (tzv. ispravljanje pogrešaka), kao i korpusi gdje su pogreške samo označene uz implicitnu normalizaciju (tzv. označavanje ispravki, označavanje pogrešaka oznakama iz unaprijed definirane taksonomije pogrešaka).

3. Učenički korpus CroLTeC

CroLTeC (Mikelić Preradović, Berać i Boras 2015: 107–126) je učenički korpus hrvatskog kao inog jezika koji se razvija na Katedri za obradu prirodnog jezika, leksikografiju i enciklopediku Odsjeka za informacijske i komunikacijske znanosti Filozofskog fakulteta Sveučilišta u Zagrebu. Javno je dostupan i pretraživ na poveznici http://nlp.ffzg.hr/resources/corpora/croltec.

Korpus sadržava eseje 755 učenika s 36 različitih materinskih jezika, među kojima su najzastupljeniji španjolski, engleski, njemački, poljski, kineski, francuski i arapski. Sastoji se od 7213 skeniranih dokumenata (originalnih učeničkih tekstova i istih tih tekstova s lektorskim ispravkama). Samo su originalni učenički tekstovi (njih 3530)[5] transkribirani i pretvoreni u XML format, a dodatnih 1217

[4] Ispravljanje pogrešnog teksta zahtijeva uspostavljanje jedne ili više ciljnih hipoteza o namjeri autora teksta i njegovu izražavanju te namjere (Lüdeling 2015: 142–145).

[5] Neki tekstovi nisu se mogli transkribirati zbog nečitljivog rukopisa ili loše kvalitete skenirane slike, a za neke tekstove postoji samo dokument s lektorskim ispravkama, ne i originalni učenički tekst.

tekstova prikupljenih u izvorno digitalnom obliku u XML format pretvoreno je naknadno. CroLTeC ukupno sadržava 4747 XML datoteka, tj. 1 054 287 pojavnica. Tekstovi su prikupljani na svih šest razina Zajedničkog europskog referentnog okvira za jezike (ZEROJ-a)[6] u Croaticumu – Centru za hrvatski kao drugi i strani jezik na Filozofskom fakultetu u Zagrebu. U okruženju TEITOK (Janssen 2016: 4037–4043) cijeli je korpus označen na više razina (lematizacija i morfosintaktičko označavanje), dok je u pilot-fazi označavanje pogrešaka provedeno na 1150 tekstova (24 % od ukupnog broja tekstova u korpusu).

U usporedbi s ostalim učeničkim korpusima slavenskih jezika: ruskim RLC-om (koji sadržava 3800 tekstova 19 nasljednih govornika ruskoga i 17 američkih govornika koji su počeli učiti ruski u odrasloj dobi), slovenskim PiKUST-om (koji sadržava 128 tekstova, od kojih su otprilike 70 % tekstova napisali hrvatski ili srpski govornici) i češkim CzeSL-om (koji sadržava 1150 tekstova stranih govornika, 540 tekstova romske djece i 490 diplomskih/magistarskih radova stranih studenata), CroLTeC svojim opsegom od 4747 tekstova koje su napisali govornici čak 36 različitih materinskih jezika predstavlja značajan leksički resurs.

Svi tekstovi obogaćeni su metapodatcima (slika 1.) o naslovu, broju i vrsti teksta te okolnostima pod kojima su nastali (domaća zadaća, dio ispita ili terenske nastave i sl.).

Task ID	1A_18
Task Title	My workday
Genre of the task	essay
Extent of the task	80+ words
Type of task	homework - weekend exercise
Description of the task	use of the direct object
Academic semester	-iti/-im
Date of the task	

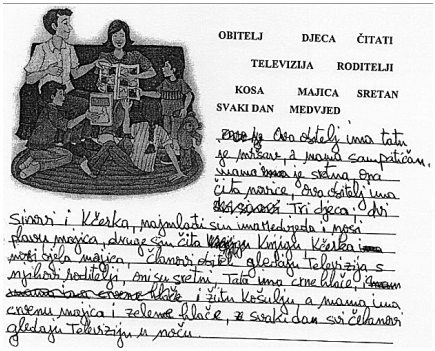
Slika 1. Zaglavlje XML dokumenta s metapodatcima

Također, korpus je moguće pretraživati prema godini rođenja, spolu, razini učenja hrvatskog jezika te prema materinskim jezicima učenika.

Sve promjene koje su napravili sami učenici (dodavanja, brisanja, transpozicije segmenata, tj. premještanja teksta s jednog mjesta u rečenici na drugo) također

[6] Engl. *Common European Framework of Reference for Languages*, *CEFR*.

su označene u transkriptima. Primjerice, na slici 2. lijevo je prikazan skeniran učenički tekst, dok je desno njegov prijepis u kojem su vidljiva brisanja (precrtani sivi tekst), tj. samoispravci.



Obitelj na slici

~~ove je~~ Ova obitelj ima tatu je mršav, a mama simpatičan, mama je sretna, ona čita novice, ova obitelj ima ~~dvi sanovi~~ Tri djeca; dvi sinovi i Kćerka, najmlađi sin ima Medvjeda i nosi plavu majica, druge sin čita ~~Knjigu~~ Knjigu, Kćerka ~~ima~~ nosi bjela majica, članovi obiteli gledaju Televizija s njihovi roditelji, oni su sretni, Tata ima crne hlače, ~~imam mama ima crvene hlače~~ i žutu Košulju, a mama ima crvenu majica i zelene hlače, z svaki dan svi čelanovi gledaju Televiziju u noču.

Slika 2. Skeniran tekst i njegov transkript

Svi su tekstovi tokenizirani, lematizirani i provedeno je morfosintaktičko označavanje riječi s pomoću označivača ReLDI (Ljubešić i Erjavec 2016: 1527–1531). Konačno, sve su ove dodatne informacije pohranjene s izvornim tekstovima u XML datotekama koje se mogu pretraživati regularnim jezikom za pretraživanje korpusa (engl. *Corpus Query Language – CQL*).

4. Označavanje pogrešaka u korpusu CroLTeC

Početna analiza malog uzorka tekstova iz neoznačenog učeničkog korpusa CroLTeC otkrila je da neizvorni govornici često rade nekoliko pogrešaka u istoj riječi (npr. pravopisnu i morfološku pogrešku) te da bi shema označavanja trebala biti u mogućnosti istovremeno prikazati sve vrste pogrešaka.

Označavanje je pogrešaka u svakom tekstu u CroLTeC-u stoga višeslojno, a shema se sastoji od triju slojeva. Pretpostavljena je ciljna hipoteza (Meurers 2015: 537–566), pri čemu je referentni jezični sustav ciljni izvorni jezik (hrvatski). Prvi sloj sadržava anonimizirani prijepis učeničkih tekstova u kojem su sačuvane izvorne učeničke pogreške, kao i samoispravci. Drugi sloj sadržava normaliziran tekst (ispravljene pogreške na razini riječi) s označenim kategorijama pogrešaka. Anotatori su obilježili pravopisne, leksičke i gramatičke pogreške

(v. tablice 1. – 3.). U trećem su sloju označene i ispravljene višerječne pogreške (tablica 5.). Konačni je rezultat procesa normaliziran, gramatički i sintaktički ispravan tekst s označenim pogreškama.

5. Kategorizacija pogrešaka

Izrada učeničkoga korpusa s označenim pogreškama za hrvatski jezik izazovan je zadatak. Inicijativa za standardizacijom i problem standardizacije taksonomija za označavanje pogrešaka kojom se želi ostvariti interoperabilnost učeničkih korpusa detaljno su izloženi u radu Stemlea i dr. (2019), gdje autori ističu da je dosad prilagođeno tek nekoliko postojećih taksonomija za označavanje novih učeničkih korpusa. Iako bi shema označavanja pogrešaka trebala biti dovoljno informativna i proširiva, također bi trebala biti lako primjenjiva i jednostavna, tj. ne bi trebala biti preopširna. Shema označavanja pogrešaka u CroLTeC-u djelomično se temelji na shemi slovenskoga razvojnog korpusa Šolar (Rozman i dr. 2013) i shemi kodiranja pogrešaka u engleskom učeničkom korpusu Cambridge Learner (Nicholls 2003: 572–581; Tono 2003: 800–809; Dagneaux i dr. 2005) te je prilagođena hrvatskom jeziku tako da je klasifikacija višerječnih pogrešaka utemeljena na taksonomiji korpusa Šolar, a iz taksonomije engleskog učeničkog korpusa preuzeto je dvoslovčano kodiranje gdje prvo slovo predstavlja opći tip pogreške (M/R/U = engl. *missing/replace/unnecesary* = nedostaje/zamijeni/nepotrebno) te dodatni kodovi (npr. AS, AG, CL). Te su dvije taksonomije upotrebljavane jer je Šolar jedini korpus slavenskih jezika koji ima razvijenu taksonomiju, a shema razvijena za korpus Cambridge Learner najčešće je upotrebljavana (jer su učenički korpusi u svijetu većinom korpusi engleskog kao inog jezika[7]).

Pogreške su najprije kategorizirane kao pogreške na razini riječi i višerječne pogreške (v. tablicu 1.).

[7] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Tablica 1. Ortografske pogreške

Označavanje pogrešaka na razini riječi			
Kategorija	Potkategorija	Kôd pogreške	Primjer
Ortografska pogreška	Nedostaje veliko slovo	MC	**z**agreb
	Nepotrebno veliko slovo	UC	Filozofski **F**akultet
	Nedostaje razmak	MB	nemogu
	Nepotreban razmak	UB	je sam
	Nedostaje slovo	MS	tjelo uvjek
	Potrebna zamjena slova	RS	ku**č**a, pokriva**ć** Upisala sam se na **economski** fakultet.
	Nepotrebno slovo	US	dragoc**i**jeno engleski**j**
	Nedostaje interpunkcija	MP	Ja ne pijem alkohol ali volim piti sokove.
	Potrebna zamjena interpunkcije	RP	Od dana kad se čovjek rodi mora putovati, pronaći svoje mjesto u svijetu, sezati više i više da bi bio zadovoljan samim sobom**,** Od dana rođenja moramo proći brojne prepreke.
	Nepotrebna interpunkcija	UP	Prošli vikend**.** nisam učila ništa novo

Pogreške na razini riječi dalje su klasificirane kao ortografske, leksičke i gramatičke pogreške. Ortografske su pogreške (tablica 1.) npr. nepravilna uporaba velikog i malog slova, interpunkcijskih znakova, metateze, pogreške na granici morfema („zatošto"), izostavljanje slova zbog nepoznavanja leksika („krupir", „Britnija", „dvje"), nepotrebna slova („duruženje", „Egipćanini", „Medvjedgrad") i neodgovarajuća uporaba dijakritičkih znakova („širomašan").

U leksičkim pogreškama leksemi nedostaju ili su suvišni, javljaju se nepostojeći leksemi ili postojeći leksemi koji izvornim govornicima nisu najprirodniji izbor (tablica 2.).

Tablica 2. Leksičke pogreške

Označavanje pogrešaka na razini riječi			
Kategorija	Potkategorija	Kôd pogreške	Primjer
Leksička pogreška	Leksem treba zamjenu, postoji u jeziku	RE	Ponekad trebam govoriti nešto što nije lijepo. Prije četiri godine sam pokvarila nogu i tri mjeseca sam ...
	Leksem treba zamjenu, ne postoji u jeziku	RN[8]	Ja imam **odnu** bratu i sestru. To je **proljeto**...
	Nedostaje leksem	ML	Moja obitelj isto [] iz Egipta. Ona kao [] ja voli trčati.
	Nepotreban (suvišan) leksem	UL	Ja živim u Zagrebu zato što [Ja] [često] želio sam učiti hrvatski jezik. [Na] vikendom idemo u klubove, u kino, u park.

[8] Sve riječi u kojima nedostaje ili je višak ili je potrebna zamjena više od jednog slova ovdje se smatraju leksičkim, a ne ortografskim pogreškama.

Tablica 3. Gramatičke pogreške

Označavanje pogrešaka na razini riječi			
Kategorija	Potkategorija	Kôd pogreške	Primjer
Gramatička pogreška	Pogrešna struktura glagolskih dopuna (višak, manjak ili pogrešna dopuna)	AS	...igrala sam [] sa sestrom. To već ovisi od regije. Seka se je plakala. Kava pijem s mlijekom.
	Pogrešna kongruencija	AG	Njihove djeca jako **voli** čitati. **Naše** sestra je bila mala. Moj tjedan je malo **dosadna**.
	Pogrešan vid glagola te lični (vrijeme, način) i nelični oblici (infinitiv, glagolski pridjev i prilog)	VT	U ožujku može **pada** snijeg i također kiša. A: Što radite? B: Hodamo gradom. A: **Dođete** i **pijete** kavu! Tjedan ranije **bismo** bili s grupom na razgledavanju grada Zagreba. Ako netko nazove da **dođe**, onda se brzo nešto kuha.
	Kolokacijska pogreška	CL	Jedem samo **biološku** hranu. Ne volim puno **crveno** vino.
	Netočna fleksija[9]	GF	... u Brazilu s **ljubavom** moga života. Možda bih **pišela** knjigu. Često **kupovaju** organsku hranu.

Gramatičke pogreške očituju se u pogrešnoj strukturi glagolskih dopuna (nedostaje obvezna dopuna ili je dopuna višak ili je pak dopuna u pogrešnom obliku). To su također pogreške u kongruenciji, ličnim i neličnim glagolskim oblicima, fleksiji ili kolokacijske pogreške (tablica 3.).

Normalizacija može od gramatički ispravnog teksta napraviti gramatički neispravan tekst, pa se naknadne korekcije označavaju kao sekundarne pogreške. Na primjer, ako se tijekom ispravljanja pogrešaka glava imeničke sintagme u pogrešnom padežu zamijeni imenicom u ispravnom padežu, mora se modificirati

[9] Fleksija se odnosi na sve promjenjive riječi koje se mijenjaju fleksijom – riječ je o pogrešci nastaloj pri sprezanju ili sklonidbi.

i pridjevski oblik da bi se u konačnici dobio normalizirani tekst (tablica 4.). Ako je učenik izvorno stavio pridjev koji se s imenicom slagao u rodu, broju i padežu, ali je imenicu napisao u pogrešnom padežu, rodu ili broju, izvorni pridjevski oblik ne može se smatrati učeničkom pogreškom.

Tablica 4. Sekundarne pogreške

Sekundarna pogreška	Pogreška uzrokovana gramatičkom kongruencijom	SEC		Kupila sam nova majica nalazim se negdje u austrijskoj Alpi.

Višerječne pogreške (tablica 5.) klasificiraju se kao gramatičke (npr. pogrešan redoslijed riječi), leksičke (npr. izostavljeni ili suvišni izrazi) i semantičke pogreške (npr. nepostojeći izrazi ili izrazi koji postoje u jeziku, ali nisu najprirodniji izbor izvornoga govornika).

Tablica 5. Višerječne pogreške

Označavanje pogrešaka na višerječnoj razini (samostalno označavanje)			
Kategorija	Potkategorija	Kôd pogreške	Primjer
Gramatika	Pogrešan redoslijed riječi	WO	Kad sam igrala sport, sam uvijek bila u najgoroj grupi.
Leksik	Nedostaje izraz	MPH	Sviđa mi se vidjeti ljude [] na moru igraju.
	Nepotreban (suvišan) izraz	UPH	**Ja sam umoran od** biti lijen.[10]
Semantika	Izraz treba zamjenu, postoji u jeziku	REP	Vaše glavno jelo ***u samo malo.***[11]
	Izraz treba zamjenu, ne postoji u jeziku	RNP	*Ne znam zašto je* ***kao da je***.[12] ***S površine****,* to je nemoguće.[13]

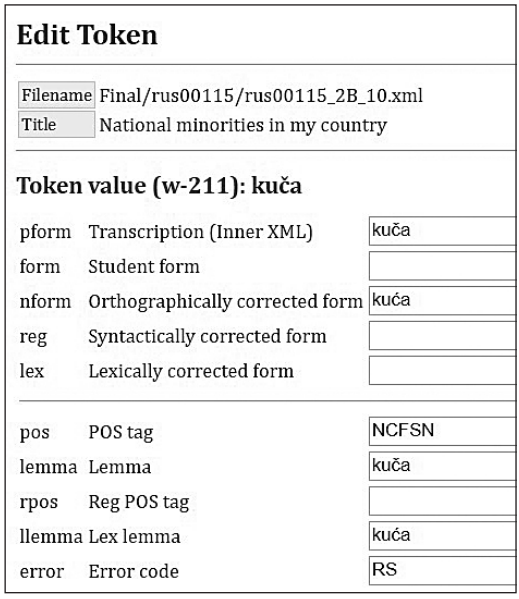
6. Označavanje pogrešaka u TEITOK-u

Označavanje pogrešaka učeničkih tekstova napravljeno je u TEITOK-u, mrežno utemeljenom sustavu za pregledavanje, stvaranje i uređivanje korpusa te tekstno

[10] Ispravljeno kao: *Umara me biti lijen.*
[11] Ispravljeno kao: *Vaše glavno jelo doći će za čas.*
[12] Ispravljeno kao: *Ne znam zašto je to tako.*
[13] Ispravljeno kao: *To je naizgled nemoguće.*

i jezično označavanje (Janssen 2016: 4037–4043). Isti se sustav upotrebljava za označavanje portugalskog učeničkog korpusa COPLE2 (del Río i Mendes 2018: 225–239).

Ovisno o problemu, tj. problemima koji utječu na izvorni učenički oblik, anotator mora za svaku pogrešnu riječ odabrati jednu ili više vrsta pogrešaka (ortografsku, leksičku, gramatičku, sekundarnu) i upisati ispravni oblik (oblike) riječi te morfosintaktičku (MSD) oznaku i lemu. Ispravljene morfosintaktičke oznake kompatibilne su s MULTEXT-East 4 sustavom oznaka za hrvatski jezik.[14] Višestruke kategorije pogrešaka mogu se navesti za danu riječ odjednom. Na primjer, kada učenički oblik sadržava ortografsku i gramatičku pogrešku (*Živim u **kuča***.).



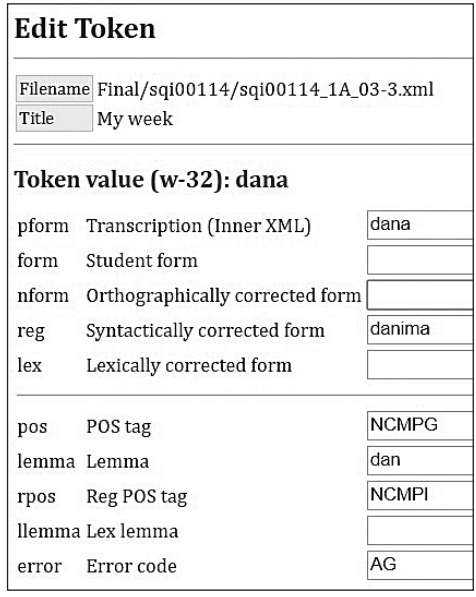
Slika 3. Označavanje pravopisne pogreške

Ako u učeničkoj rečenici postoji pravopisna pogreška, u TEITOK-u se upisuje ortografski ispravan oblik (*nform*), kao i odgovarajuća morfosintaktička oznaka (*pos*) te lema (*llemma*), ukoliko je to potrebno, kao što je prikazano na slici 3.: učenik je napisao „kuča" umjesto „kuća", pa se pravopisno ispravni oblik unosi pod *nform*, morfosintaktička oznaka (*pos*) ne treba ispravljanje (*rpos* stoga osta-

[14] http://nl.ijs.si/ME/V4/

je prazan), dok je ispravna lema (*llemma*) različita od učeničke leme (*lemma)* i stoga treba biti unesena u odgovarajuću kućicu.

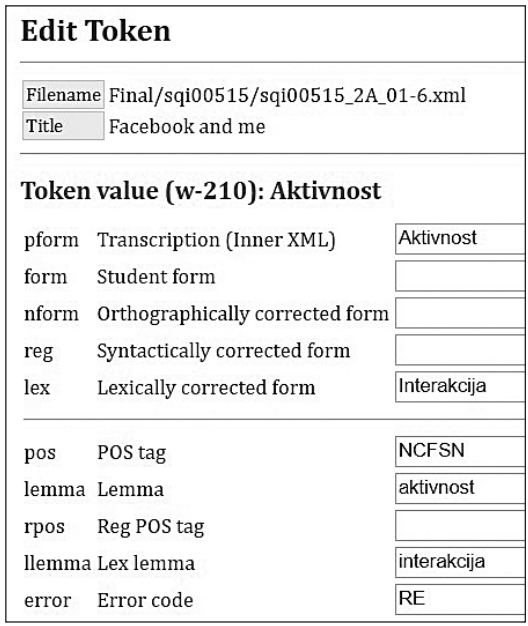
Slika 4. prikazuje primjer gramatičke pogreške: riječ koju upotrebljava učenik generira negramatički izraz. Učenik je napisao „dana" umjesto „danima" u rečenici „Na tečaj jezika idem *radnim danima*." i stoga postoji pogreška u kongruenciji označena kao AG. U TEITOK se unosi ispravljen oblik (*reg*) kao i odgovarajuća morfosintaktička oznaka (*rpos*).



Slika 4. Označavanje gramatičke pogreške

Konačno, ako postoji leksička pogreška u učenikovoj rečenici, tj. riječ je gramatički ispravna, ali nije najprirodniji izbor izvornoga govornika, unosi se leksički ispravan oblik (*lex*), kao i odgovarajuća lema (*llema*) i morfosintaktička oznaka (*rpos*) ako je to potrebno, kao što je prikazano na slici 5.: učenik je napisao „aktivnost" u kontekstu gdje je „interakcija" logičniji izbor („*Aktivnost* na fejsu nije prava aktivnost i ponekad mislimo da imamo puno prijatelja.").

Na slici 5. ispravljena je samo lema (*llema*) i unesen kôd pogreške (RE) jer se ispravna lema razlikuje od leme oblika u učeničkoj pogrešci, a morfosintaktička oznaka *rpos* ima istu vrijednost kao i morfosintaktička oznaka *pos* i stoga ta kućica ostaje prazna.



Slika 5. Označavanje leksičke pogreške

7. Rezultati označavanja pogrešaka

Do sada je u pilot-fazi označeno 1150 tekstova na svim trima slojevima (24 % od ukupnog broja tekstova u korpusu). Označeno je 29 150 pogrešaka, od kojih je većina pogrešaka na razini riječi (27 776 ili 95 % oznaka). Oznake na razini riječi imaju sljedeću distribuciju: 8393 pravopisnih pogrešaka, 7683 leksičkih, 10 848 gramatičkih pogrešaka i 842 sekundarne pogreške, koje nisu učeničke pogreške, nego naknadne pogreške nastale zbog kongruencije. Analiza pokazuje da su anotatori, unatoč tomu što su prošli obuku o taksonomiji pogrešaka, tijekom označavanja nailazili na probleme a) interferencije, b) interpretacije, c) reda riječi i d) kolokvijalizama.

A) interferencija

Budući da su anotatori bili obučeni studenti informacijskih znanosti, a ne stručnjaci u području usvajanja inog jezika, od njih se nije očekivalo uočavanje slučajeva jezične interferencije materinskog jezika inojezičnih

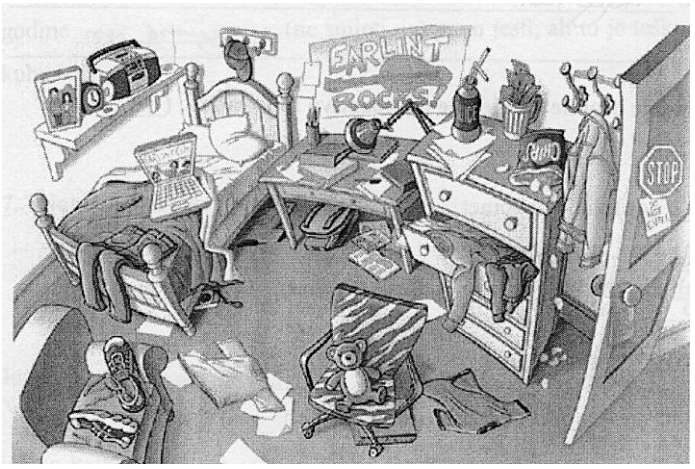
govornika. Rečenica poput „Maja ima savršeno surfersko prkno.” gramatički je ispravna, ali je njezin autor, izvorni govornik češkog jezika, bio zaveden tzv. lažnim srodnicima i pretpostavio da je „prkno” (hrv. vulg. „stražnjica”) hrvatski ekvivalent češke imenice „prkno” (daska) u „surfařské prkno” (daska za jedrenje).

B) interpretacija

Kod nekih vrsta pogrešaka anotatorima je bio problem definirati granice interpretacije. Rečenica „Često i se bavim sportom u sveučilištu jer ima velike igralište, i ja sam udružio tim košarke.” gramatički je netočna, ali se otprilike može interpretirati kao „Pridružio sam se košarkaškom timu.”. U takvim je slučajevima zadaća anotatora interpretacija, a ne korekcija. Rečenica se može ispraviti u „Često se bavim sportom na Sveučilištu jer ima veliko igralište i ja sam se pridružio košarkaškom timu.”, ili „Često se bavim sportom na Sveučilištu jer ima veliko igralište i ja sam osnovao košarkaški tim.”, pri čemu je potonja manje vjerojatna, ali bliža izvornoj rečenici. Ovdje je anotatorima teško dati jasne smjernice.

C) red riječi

Red riječi u rečenici u hrvatskom jeziku odražava informacijsku strukturu i anotatoru može biti teško odlučiti (čak i kad vidi kontekst, kao što je prikazano na slici 6.) je li pogreška prisutna. Rečenica „Torba je i na podu.” redoslijedom riječi sugerira da postoji torba na više od jednoga mjesta u sobi, iako je vjerojatnije tumačenje da se na podu, između ostaloga, nalazi i torba. Posljednje tumačenje pak zahtijeva drugačiji redoslijed riječi: „Na podu je i torba”.



Slika 6. Učenici u eseju opisuju sobu sa slike

D) kolokvijalizmi

Neizvorni govornici često upotrebljavaju kolokvijalizme jer ih čuju u svakodnevnoj neformalnoj komunikaciji s izvornim govornicima (npr. „Pazim kaj jedem.”). Čak i kad su ti izrazi gramatički ispravni, anotatori ih trebaju zamijeniti izrazima standardnog jezika, što također može biti teško jer ti izrazi prosječnom govorniku zvuče vrlo prirodno.

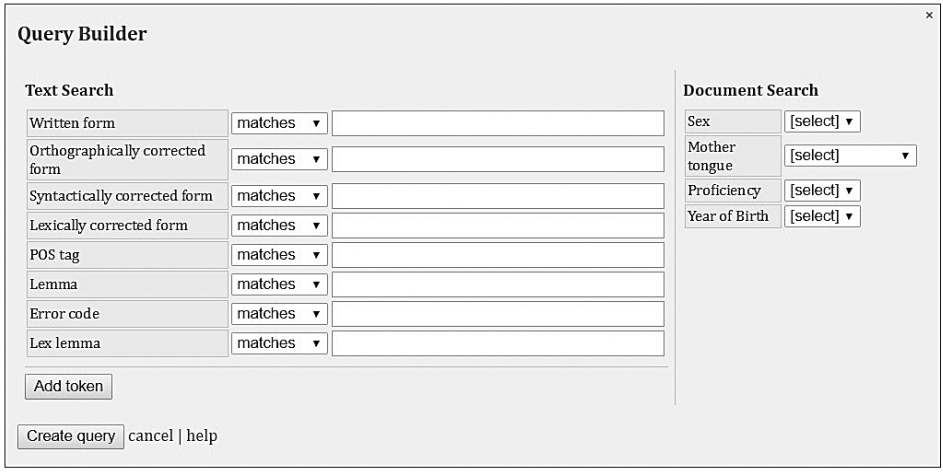
Naposljetku, anotatori su tijekom rada na označavanju izrazili dvojbe i neslaganja oko vrste pogreške (npr. *Taj je koncept nov u* ***Hrvatskom***., gdje pogreška može biti označena kao ortografska – *hrvatskom* ili gramatička – *Hrvatskoj*).

Također se pokazalo da analiza učeničkih tekstova koje su ispravili lektori hrvatskog kao inog jezika anotatorima nije bila korisna. Lektori zbog pedagoških razloga često ispravljaju samo one pogreške koje odražavaju neadekvatno usvajanje jezičnoga materijala na specifičnoj razini učenja jezika, što je u stvarnosti rezultiralo vrlo malim brojem označenih i ispravljenih pogrešaka na kojima su anotatori mogli temeljiti svoj izbor.

Zbog svih gore navedenih poteškoća, nakon dovršetka pilot-faze jezični će stručnjaci ekstrahirati pogreške iz korpusa prema oznakama te, po potrebi, ispravljati pogrešne oznake.

8. Pretraživanje korpusa CroLTeC

U TEITOK-u je razvijeno grafičko sučelje za pretraživanje kako bi se korisnicima omogućilo jednostavno oblikovanje upita za pretraživanje korpusa. Sučelje omogućuje odabir najrelevantnijih kategorija polja za pretraživanje (ortografski/leksički/gramatički ispravljeni oblici, morfosintaktička oznaka, izvorna lema, ispravljena lema) te sadržava popis pretraživih metapodataka koji se mogu upotrebljavati kao filtri preko padajućeg izbornika (spol neizvornog govornika, materinski jezik, razina učenja jezika i godina rođenja). Korisnici ne moraju unaprijed znati koje podatke sadržava korpus, a sučelje im ipak omogućuje pristup tim informacijama. Budući da se korpusni alat koristi jezikom CQL, TEITOK prevodi upite korisnika na taj jezik. Slika 7. prikazuje sastavljanje upita u CroLTeC-u s dostupnim poljima za pretraživanje i mogućim vrijednostima metapodataka.



Slika 7. Sastavljanje upita u CroLTeC-u

Korisničko sučelje omogućuje i postavljanje upita u CQL jeziku te ekstrakciju kolokacija. Sučelje TEITOK pretvara upite korisnika koji nisu upoznati s CQL-om u taj regularni jezik. Primjerice, ako korisnik na padajućem izborniku odabere da lema počinje nizom *jel*, upit će se automatski pretvoriti u *[lemma = " jel. * "] within text.*

Korisnik može upisati bilo koju hrvatsku riječ kao upit i istražiti sve moguće kolokate s njihovim morfosintaktičkim oznakama te dobiti uvid u različite susjedne kolokate.

Korisnici mogu pretraživati učestalost različitih vrsta pogrešaka i dobiti konkordancije za određenu vrstu pogreške, a mogu dobiti i čestotnu raspodjelu tipova pogrešaka za cijeli korpus.

Korisnicima koji u korpusu traže samo primjere sučelje nudi mogućnost pretraživanja na način sličan uobičajenim tražilicama. Bilo koji upit koji sadržava samo jednostavne znakove tumači se kao pretraživanje s pomoću tražilice, što znači da korisnik može upisati „majka" za pretraživanje rečenica koje sadržavaju „majka", kao što bi to učinio i npr. na Googleu. No korisnik može upisati *["maj.*"]* i pronaći sve riječi koji počinju nizom „maj", kao što su „majica", „majka", „majmun", „majoneza", „major" i „majstor" u svim padežnim oblicima.

9. Zaključak

Shema za označavanje i označene pogreške u učeničkom korpusu sa svim ostalim informacijama (metapodatcima i učeničkim samoispravcima) mogu biti korisni znanstvenicima koji istražuju usvajanje inog jezika (MacWhinney 2017: 254–275) ili računalno potpomognuto učenje jezika (Meurers i Dickinson 2017: 66–95).

Predložena shema za označavanje pogrešaka može se upotrijebiti u nekom budućem razvojnom korpusu hrvatskog jezika (primjerice, korpusu izvornih govornika – školaraca), no s obzirom na relativno dobru primjenjivost sheme označavanja slovenskog razvojnog korpusa i sheme označavanja engleskog učeničkog korpusa na hrvatski jezik također se može upotrijebiti i za druge jezike koji tek razvijaju svoje učeničke korpuse (pri čemu će se za slavenske jezike vjerojatno moći upotrebljavati sve potkategorije, dok bi se za ostale jezike za označavanje mogle upotrijebiti barem glavne kategorije).

CroLTeC s označenim pogreškama trebao bi omogućiti istraživačima i stručnjacima koji se bave proučavanjem hrvatskog kao inog jezika da istraže me-

đujezičnost, uoče koji aspekti jezika određenim skupinama učenika uzrokuju najviše poteškoća u usvajanju hrvatskoga i prilagode nastavne materijale (ne samo prema razini učenja hrvatskog jezika nego i prema materinskom jeziku učenika). Konačno, korpus s označenim pogreškama trebao bi poslužiti i kao polazna točka za izradu računalnih alata za ispravljanje leksičkih pogrešaka, odstupanja u glagolskim vremenima, kongruenciji i kolokacijama te za inovacije u nastavi hrvatskog kao inog jezika.

Literatura

Dagneaux, Estelle; Granger, Sylviane; Meunier, Fanny; Thewissen, Jennifer; Denness, Sharon i dr. 2005. *Error Tagging Manual. Version 1.2.* Centre for English Corpus Linguistics. Université Catholique de Louvain. https://hdl.handle.net/2078.1/75592 (pristupljeno 10. veljače 2020.).

Del Río, Iria; Mendes, Amália. 2018. Error annotation in the COPLE2 corpus. *Revista Da Associação Portuguesa De Linguística* 4. 225–239.

Díaz-Negrillo, Ana i dr. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum* 36/1–2. 139–154.

Dobric, Nikola. 2015. Quality measurements of error annotation – Ensuring validity through reliability. *The European English Messenger* 24/1. 36–42.

Flor, Michael; Futagi, Yoko. 2012. On using context for automatic correction of non-word misspellings in student essays. *Proceedings of the seventh workshop on building educational applications Using NLP.* Ur. Tetreault, Joel; Burstein, Jill; Leacock, Claudia. Association for Computational Linguistics. 105–115.

Gaillat, Thomas, Sébillot, Pascale; Ballier, Nicolas. 2014. Automated classification of unexpected uses of this and that in a learner corpus of English. *Recent Advances in Corpus Linguistics: Developing and Exploiting Corpora* 78. Ur. Vandelanotte, Lieven; Davidse, Kristin; Gentens, Caroline. 309–324. doi.org/10.1163/9789401211130_015.

Granger, Sylviane. 2003. Error-tagged learner corpora and CALL: A promising synergy. *CALICO journal* 20. 465–480.

Granger, Sylviane. 2008. Learner corpora. *Corpus Linguistics. An International Handbook.* Ur. Lüdeling, Anke; Kytö, Merja. Walter De Gruyter. Berlin – New York. 259–274.

Hana, Jirka; Rosen, Alexandr; Škodová, Svatava; Škodová, Barbora. 2010. Error-tagged learner corpus of Czech. *Proceedings of the Fourth Linguistic Annotation Workshop.* Ur. Xue, Nianwen; Poesio, Massimo. Association for Computational Linguistics.

Uppsala. 11–19. https://dl.acm.org/doi/10.5555/1868720.1868722 (pristupljeno 10. veljače 2020.).

Janssen, Maarten. 2016. TEITOK: Text-Faithful Annotated Corpora. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016*). Ur. Calzolari, Nicoletta; Choukri, Khalid; Declerck Thierry; Goggi, Sara; Grobelnik, Marko; Maegaard, Bente; Mariani, Joseph; Mazo, Helene; Moreno, Asuncion; Odijk, Jan; Piperidis, Stelios. European Language Resources Association (ELRA). Pariz. 4037–4043.

Ljubešić, Nikola; Erjavec, Tomaž. 2016. Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016*). Ur. Calzolari, Nicoletta; Choukri, Khalid; Declerck Thierry; Goggi, Sara; Grobelnik, Marko; Maegaard, Bente; Mariani, Joseph; Mazo, Helene; Moreno, Asuncion; Odijk, Jan; Piperidis, Stelios. European Language Resources Association (ELRA). Pariz. 1527–1531.

Lüdeling, Anke; Hirschmann, Hagen. 2015. Error annotation systems. *The Cambridge Handbook of Learner Corpus Research*. Ur. Granger, Sylvaine; Gilquin, Gaëtanelle; Meunier, Fanny. Cambridge University Press. 135–158.

MacWhinney, Brian. 2017. A Shared Platform for Studying Second Language Acquisition. Conceptual Review Article. *Language Learning* 67/S1. 254–275. doi.org/10.1111/lang.12220.

Meurers, Detmar; Dickinson, Markus. 2017. Evidence and Interpretation in Language Learning Research: Opportunities for Collaboration with Computational Linguistics. Conceptual Review Article. *Language Learning* 67/S1. 66–95. doi.org/10.1111/lang.12233.

Meurers, Detmar. 2015. Learner Corpora and Natural Language Processing. *The Cambridge Handbook of Learner Corpus Research*. Ur. Granger, Sylvaine; Gilquin, Gaëtanelle; Meunier, Fanny. Cambridge University Press. 537–566.

Mikelić Preradović, Nives; Berać, Monika; Boras, Damir. 2015. Learner Corpus of Croatian as a Second and Foreign Language. *Multidisciplinary Approaches to Multilingualism*. Ur. Cergol Kovačević, Kristina; Udier, Sanda Lucija. Peter Lang. Frankfurt am Main. 107–126.

Nicholls, Diane 2003. The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. *Proceedings of the Corpus Linguistics 2003 Conference*. Ur. Archer, Dawn i dr. Lancaster University. 572–581.

Rakhilina, Ekaterina i dr. 2016. Building a learner corpus for Russian. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC 2016*. Ur. Volodina, Elena i dr. Linköping LiU Electronic Press. 66–75.

Rastelli, Stefano. 2009. Learner corpora without error tagging. *Linguistik Online* 38/2. 57–66. doi.org/10.13092/lo.38.507.

Rosén, Alexander. 2017. Introducing a corpus of non-native Czech with automatic annotation. *Language, Corpora and Cognition*. Ur. Pęzik, Piotr; Waliński, Jacek. Peter Lang. Bern – Varšava. 163–180.

Rosén, Dan; Wirén, Mats; Volodina, Elena. 2018. Error coding of second language learner texts based on mostly automatic alignment of parallel corpora. *Proceedings of the CLARIN Annual Conference 2018.* Ur. Skadina, Inguna; Eskevich, Maria. Pisa. 181–184.

Rozman, Tadeja; Stritar Kučuk, Mojca; Kosem, Iztok; Krek, Simon; Krapš Vodopivec, Irena; Arhar Holdt, Špela; Stabej, Marko. 2013. *Learners' corpus Šolar 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1036 (pristupljeno 10. veljače 2020.).

Selinker, Larry. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching* 10/1–4. 209–232. doi.org/10.1515/iral.1972.10.1-4.209.

Stemle, Egon i dr. 2019. Working together towards an ideal infrastructure for language learner corpora. *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*. Ur. Abel, Andrea; Glaznieks, Aivars; Lyding, Verena; Nicolas, Lionel. Presses universitaires de Louvain, Louvain-la-Neuve. 440–478.

Stritar, Mojca. 2009. Slovene as a foreign language: The pilot learner corpus perspective. *Slovenski jezik – Slovene Linguistic Studies* 7. 135–152.

Tono, Yukio. 2003. Learner corpora: Design, development and applications. *Proceedings of the Corpus Linguistics 2003 Conference*. Ur. Archer, Dawn; Rayson, Paul; Wilson, Andrew; McEnery, Tony. Lancaster University Centre for Computer Corpus Research on Language. 800–809.

Error-Tagging of CroLTeC (Electronic Learner Corpus of Croatian as a Foreign Language)

Abstract

The paper describes the error-tagging scheme developed for the CroLTeC learner corpus (http://nlp.ffzg.hr/resources/corpora/croltec/) – the first electronic learner corpus of Croatian as a foreign language. CroLTeC contains essays collected from 755 students with 36 different mother tongues, among which the most prominent were Spanish, English, German, Polish, Chinese, French, and Arabic. It consists of 4,747 essays, out of which 1,217 were digitally born, while 3530 essays were scanned, transcribed in RTF format,

and converted into XML format. CroLTeC has a total of 1,054,287 tokens, and essays have been collected on all 6 levels of Common European Framework of Reference for Languages (CEFR) at Croaticum – Center for Croatian as Second and Foreign Language at the Faculty of Humanities and Social Sciences in Zagreb, Department of Information Sciences, Natural Language Processing group. All CroLTeC essays contain metadata about the title, number, and type of essay (homework, part of an exam or field class, etc.). Data were lemmatized and annotated with morphosyntactic tags with the ReLDI tagger (Ljubešić et al., 2016). Also, the corpus is searchable by age, sex, language proficiency level, and the mother tongue of the learner.

The error-tagging scheme is partially based on Šolar (the scheme of Developmental corpus of Slovene) and the error-coding of the Cambridge Learner Corpus and further tailored to the Croatian language. The goal of the development of the error-tagging scheme is to build a sub-corpus that will serve as a repository of authentic data about the learner's interlanguage. It should enable researchers and teachers of Croatian as a foreign language to explore the interlanguage, to discover the aspects of the grammar that are the most difficult to master and to tailor teaching materials to different groups of learners (not only according to their Croatian language proficiency level but also to their first language). Finally, the error-tagged sub-corpus should also serve as a starting point for designing computer-aided tools to correct lexical errors, misuse of verbal tenses, phrasal verbs, and collocations.

Ključne riječi: učenički korpusi, CroLTeC, obilježavanje pogrešaka, ispravljanje pogrešaka, normalizacija

Keywords: learner corpora, CroLTeC, error annotation, error-tagging, normalization