Anotacija „ITALSERB“ govornog učeničkog korpusa

Predmet ovog izlaganja je anotacija korpusa „ITALSERB“, korpusa italijanskog kao L2 izrađenog na Katedri za italijanski jezik i književnost Filološkog fakulteta u Beogradu. Ovaj korpus obuhvata 25 sati u potpunosti transkribovanih audio snimaka usmene produkcije preko 170 različitih srbofonih studenata italijanskog kao L2, sa približno 200.000 tokena. Sudeći po njegovim dimenzijama, ITALSERB se može uvrstiti u velike korpuse za potrebe proučavanja usvajanja L2. S obzirom na to da korpus sadrži produkciju studenata jedne generacije tokom četiri godine studija, ali i produkciju četiri različite generacije studenata na različitim godinama studija, korpus poseduje kako transverzalni, tako i longitudinalni karakter, što ga čini veoma vrednim resursom za jezička istraživanja govorne produkcije studenata italijanskog jezika. Pored ortografske transkripcije govora, korpus sadrži i informacije o upotrebi stranih reči, dijalekatskih reči, negramatičnih oblika i lapsusa, kao i informacije o različitim vrstama pauza u govoru, prekidima unutar reči, pogrešnim startovima, varijacijama u tonu glasa, signala u funkciji potvrde prijema, preklapanja turnusa govornika, kao i drugih paralingvističkih i nelingvističkih pojava, što predstavlja izuzetnu osnovu za mnoge vrste jezičkih istraživanja.

U izlaganju bi bilo objašnjeno usklađivanje prvobitnog sistema transkripcije sa TEI smernicama i njihovo pretvaranje u XML format kompatibilan sa TXM programom za pregled i obradu korpusa. Zatim, bio bi predstavljen proces anotacije prema vrstama reči (*Part-Of-Speech tagging, POS-tagging*) pomoću programa *TreeTagger*, pri čemu je svakoj reči dodeljena oznaka vrste reči i odgovarajuća lema, što značajno proširuje spektar jezičkih istraživanja koja se mogu vršiti na ovom korpusu. Ukupno trajanje do sada anotiranih konverzacija je 3 sata i 19 minuta.

U izlaganju bi takođe bio predstavljen sistem metapodataka vezanih za govornike, tip i nivo tekstova, pomoću kojih je moguće suziti opseg analiza na određeni jezički nivo ili čak pratiti usvajanje određenih jezičkih pojava u govoru pojedinačnih učenika.

S obzirom na jedinstveni karakter ovog resursa i teškoće u usklađivanju različitih sistema i nivoa anotacije, bila bi predstavljena i metodološka i softverska rešenja koja su bila neophodna za usklađivanje prvobitne transkripcije sa TEI smernicama, automatsko numerisanje tokena i tagova, proces anotacije pomoću programa *TreeTagger*, i kreiranje jedinstvenog korpusa u programu TXM.

Na kraju, bile bi izložene određene vrste jezičkih analiza koje je moguće izvršiti na korpusu, kao i dalji koraci u procesu anotacije i omogućavanja pristupa korpusu.