SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI

AK. GOD. 2016. / 2017.

Ivan Nedić

Model izrade učeničkog korpusa

diplomski rad

Mentorica: prof. dr. sc. Nives Mikelić Preradović

Zagreb, rujan 2017.

Sadržaj

1. Uvod

Prikupljanje velikih količina jezika radi boljeg i lakšeg objašnjavanja riječi postoji barem više od stotinu godina. Isprva je takav proces bio dugotrajan, mukotrpan i prepun problema, a često gotovo i nemoguć. Mogućnosti lingvističkih analiza nad takvim tekstovima bile su ograničene jednako koliko je i valjanost analiza bila upitna. Razvoj računala i informacijskih tehnologija u moderno doba uvelike je riješio tehničke probleme i omogućio bržu i bolju izradu korpusa koji su u mogućnosti ponuditi pouzdanije lingvističke analize i reprezentativan jezik. Međutim, u vremenu globalizacije i globalne umreženosti potreba za jezičnim resursima veća je nego ikada. Nužnost gotovo svakodnevne komunikacije s osobama iz drugih dijelova svijeta koji ne govore naš materinski jezik povećala je značaj učenja stranih jezika, a samim time i potrebu za pomagalima. Područja usvajanja drugog jezika i poučavanja stranog jezika s vremenom su prihvatila korpusni pristup i korpuse kao vrijedan alat koji pruža brojne prednosti. Daljnjim razvojem korpusne lingvistike i korpusa došlo je do specijalizacije i podjele korpusa ovisno o namjeni. Glavna tema ovog rada jedna je od podvrsta korpusa – učenički korpusi. Njihova je pojava i dalje relativno nova iako sve više istraživača i lingvista iz raznih zemalja uviđa njihov značaj. U skladu s tim povećanim interesom za učeničke korpuse došlo je i do značajnog porasta broja novih učeničkih korpusa koji su sada u izradi za mnoge jezike. Prvi dio rada odnosi se na korpusnu lingvistiku i njezin položaj unutar lingvističke teorije. Osim toga definirani su korpusi u općenitom smislu i ukratko je predstavljena njihova povijest. Također su navedene i ukratko objašnjenje glavne vrste korpusa. U posljednjem dijelu rada predstavljeni su učenički korpusi tako što su definirani i navedena glavna područja njihove primjene, a navedeni su i neki gotovi i tekući projekti koji prikazuju trenutačno stanje u tom području. Glavna je tema sama izrada učeničkih korpusa te su objašnjeni neki metodološki problemi i razmatranja koje treba uzeti u obzir prilikom izrade, kao i apstraktni postupak i vrste glavnih analiza koje se na gotovom korpusu mogu napraviti, a koje su uvijek krajnji cilj same izrade učeničkog korpusa. Na samom kraju rada ukratko su predstavljena dva alata za korpuse te je opisan postupak izrade učeničkih korpusa i mogućnosti u istima, a predstavljen je i cjelokupan postupak izrade učeničkog korpusa COPLE2.

2. Korpusna lingvistika

Principi korpusne lingvistike postoje već gotovo jedno stoljeće. Leksikografi ili izrađivači rječnika skupljaju primjere jezika u uporabi da bi lakše precizno definirali riječi najmanje od kraja 19. stoljeća. Prije računala takvi primjeri jezika u principu su se prikupljali na malim papirima i organizirali prema Dirichletovom principu. Pojava računala dovela je do nastanka modernih korpusa[1]. Suvremena korpusna lingvistika zasebna je grana lingvistike koja se bavi jezičnom analizom strojno izrađenih korpusa pisanoga ili govornoga jezika.[2] Razvoju moderne korpusne lingvistike doprinijelo je mnogo poznatih znanstvenika, a među ostalima i Leech, Biber, Johansson, Francis, Hunston, Conrad i McCarthy. Iako je njihov doprinos u prošlosti (ali još i danas) bio vrlo velik, mnogi korpusni lingvisti smatraju Johna Sinclaira možda i najutjecajnijim znanstvenikom moderne korpusne lingvistike. Sinclair je otkrio da riječ sama po sebi ne prenosi značenje nego da značenje često tvori nekoliko riječi u nizu. Ta je ideja temelj korpusne lingvistike.[3]

Korpusna lingvistika pristupa istraživanju jezika u uporabi s pomoću korpusa. Moglo bi se reći da korpusna lingvistika pokušava dati odgovor na dva temeljna istraživačka pitanja:

1) Koji su uzorci povezani s leksičkim ili gramatičkim značajkama?
2) Kako se ti uzorci razlikuju među varijacijama i registrima?[4]

Korpusna lingvistika ne može dati negativne dokaze, odnosno korpus ne može dati informacije o tome što je moguće ili ispravno ili što nije moguće ili je neispravno u jeziku, nego nam može dati informacije samo o tome što je ili nije prisutno u korpusu. Bitno je napomenuti da nedostatak određenog načina izražavanja neke ideje ne znači da je korpus neispravan. Vjerojatno je da taj način jednostavno nije vrlo uobičajen u registru koji je sadržan u korpusu.[5]

Korpusna lingvistika također ne može objasniti zašto je nešto tako kako jest – može samo pokazati što jest. Da bismo otkrili zašto, mi kao korisnici jezika služimo se našom intuicijom. Korpusna lingvistika isto tako ne može ponuditi sav mogući jezik u jednom trenutku. S obzirom da korpus prema definiciji mora biti izrađen prema određenim načelima, jezik koji se nalazi u

[1] Bennett R., G. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. University of Michigan Press, 2010.
[2] Klobučar Srbić, I. Obol korpusne lingvistike suvremenoj leksikografiji. // Studia lexicographica. 2 (3) (2008)
[3] Bennett R., G. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. University of Michigan Press, 2010.
[4] Ibid.
[5] Ibid.

korpusu nije nasumičan nego planiran. Međutim, koliko god korpus bio isplaniran, velik i izrađen prema načelima, ne može biti reprezentativan za sav jezik. Drugim riječima, čak i u korpusu koji se sastoji od milijardu riječi, kao što je *Cambridge International Corpus*, svi primjeri uporabe jezika možda, odnosno sigurno nisu prisutni.[6] U suštini, korpusna lingvistika daje nam uvid u uporabu jezika danas i način na koji se taj jezik upotrebljava u različitim kontekstima te tako doprinosi u raznim jezičnim područjima, kao što je primjerice podučavanja jezika.[7]

2.1. Korpusna lingvistika i lingvistička teorija

Jedno od glavnih pitanja o korpusnoj lingvistici jest pitanje je li metodologija ili teorija? Većina korpusnih lingvista neće na to pitanje odgovoriti jednim od ponuđenih odgovora, ali prilikom analiziranja jezika s pomoću korpusa upotrebljava se određena „metoda“.[8] Tako Hunston vjeruje da je korpusna lingvistika u suštini metodologija ili skup metodologija, prije nego teorija opisa jezika. Prema njoj, korpusna lingvistika u suštini označava sljedeće:

- promatranje prirodno nastalog jezika;
- promatranje relativno velikih količina takvog jezika;
- promatranje relativnih frekvencija, bilo u neobrađenom obliku ili skroz statističke operacije;
- promatranje obrazaca povezanosti, bilo između značajke i vrste teksta ili između skupina riječi.[9]

Ako na taj način svedemo korpusnu lingvistiku na njezinu srž, ona se čini neutralna u pogledu teorije, iako prakticiranje korpusne lingvistike nikada nije neutralno s obzirom da svaka osoba definira što označava „značajku“ i koje frekvencije je potrebno promatrati, u skladu s teoretskim pristupom o tome što je u jeziku bitno. Pristupi uporabi korpusa koji se u suštini oslanjanju na postojanje kategorija dobivenih iz nekorpusnih istraživanja jezika ponekad se nazivaju **korpusno utemeljenima** (eng. *corpus-based*). Istraživanja te vrste mogu ispitati hipoteze koje proizlaze iz gramatičkih opisa temeljenih na intuiciji ili ograničenim podacima.[10]

[6] Bennett R., G. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. University of Michigan Press, 2010.
[7] Ibid.
[8] Ibid.
[9] Hunston, S. Corpus linguistics. // Encyclopedia of Language & Linguistics (Second Edition), 2006.
[10] Ibid.

Sinclair pak tvrdi da vrste obrazaca koje se mogu primijetiti u korpusu (i nigdje drugdje) zahtijevaju opise značajno drugačije vrste od onih koji su obično dostupni. Opisi i teorije koje oni potaknu smatraju se **korpusno upravljanima** (eng. *corpus-driven*). Neki od izazova za tradiciju koje uključuju korpusno upravljane teorije su sljedeći:

- Leksik i gramatika nisu odvojeni i gramatika nije apstraktni sustav u temelju jezika
- Bilo kakav izbor uvelike je ograničen izborom leksika
- Značenje nije atomističko i ne leži u riječima, nego je prozodijsko, što znači da pripada varijabilnim jedinicama značenja i uvijek se nalazi u tekstu.[11]

Podjelu istraživačkih pristupa u korpusnim istraživanjima na *djelomično korpusno utemeljen pristup* i *potpuno korpusno utemeljen pristup* uvela je Tognini-Bonelli s ciljem razlikovanja pristupa koji se korpusom služi „da bi se provjerile neke unaprijed postavljene hipoteze“ (eng. *corpus-based approach*) od pristupa koji „hipoteze postavlja isključivo na temelju rezultata korpusne analize“ (eng. *corpus-driven approach*).[12]

2.2. Korpusni pristup

Već je prije istaknuto kako je korpus temelj za svako istraživanje teksta, bez obzira na to promatra li se kao jezična građa ili nešto što se putem teksta tek ostvaruje.[13] Ono što karakterizira korpusnu lingvistiku jest svakako korpusni pristup. Za korpusni pristup možemo reći da se sastoji od četiri glavne značajke:

1) Korpusni je pristup empirijski, što znači da analizira stvarne obrasce jezika u uporabi u prirodnim tekstovima.

Ključni dio ove značajke korpusnog pristupa jest autentičan jezik. Korpusi se sastoje od udžbenika, fikcije, časopisa, akademskih tekstova, literature, novina, telefonskih razgovora, radio i tv emisija i dr. Ukratko, sve situacije iz stvarnoga svijeta u kojima se odvija bilo kakva lingvistička komunikacija mogu se nalaziti u korpusu.[14]

[11] Hunston, S. Corpus linguistics. // Encyclopedia of Language & Linguistics (Second Edition), 2006.
[12] Štrkalj Despot, K.; Möhrs, C. Pogled u e-leksikografiju. // Časopis Instituta za hrvatski jezik i jezikoslovlje. 41/2 (2015.)
[13] Klobučar Srbić, I. Obol korpusne lingvistike suvremenoj leksikografiji. // Studia lexicographica. 2 (3) (2008)
[14] Bennett R., G. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. University of Michigan Press, 2010.

2) Korpusni pristup kao temeljom za analizu koristi se velikom zbirkom prirodno nastalih tekstova prikupljenih prema određenim načelima.

Ta značajka korpusnog pristupa odnosi se na sam korpus. To znači da je moguće raditi s pisanim korpusima, govornim korpusima, akademskim govornim korpusima itd.[15]

3) Korpusni pristup uvelike upotrebljava računala za analize.

Računala ne sadrže samo korpuse, ona pomažu i u analizi jezika u korpusu. Korpusu se pristupa i korpus se analizira s pomoću raznog softvera. Ukratko, bez računala nije moguće učinkovito upotrijebiti korpuse ili primijeniti korpusni pristup.[16]

4) Korpusni pristup ovisi o kvantitativnim i kvalitativnim analitičkim tehnikama.

Ta značajka korpusnog pristupa ističe važnost naše intuicije kao stručnih korisnika nekog jezika. Kvantitativne rezultate dobivene iz korpusa tada kvalitativno analiziramo da bismo pronašli značaj. Kvalitativna analiza rezultata uključuje primjerice pregled priloga za stupnjevanje u uporabi radi razumijevanja situacija u kojima se upotrebljavaju. Na taj način dolazimo do odgovora na pitanje *zašto*?[17]

Korpusni pristup, odnosno korpusna metodologija, može se jednostavno primijeniti u mnogim lingvističkim disciplinama: fonologiji, morfologiji, sintaksi, sociolingvistici, kognitivnoj lingvistici. Filološke su znanosti po definiciji usmjerene na istraživanje tekstova, a računalnu obradu tekstualne građe mogu primjenjivati i znanost o književnosti (kada konzultira činjenice teksta), povijest (kada zahtijeva uvid u dokumente, što inače spada u pomoćne povijesne znanosti poput arhivistike), djelomično i arheologija (kada proučava na/t/pise), ali i psihologija i sociologija. Svrha računalne obrade kao istraživačkog alata, instrumenta ili pomagala, trebalo bi biti omogućavanje znanstvenicima (i svima ostalima) ususstavljenog i brzog pristupa velikim količinama teksta.[18]

[15] Bennett R., G. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. University of Michigan Press, 2010.
[16] Ibid.
[17] Ibid.
[18] Klobučar Srbić, I. Obol korpusne lingvistike suvremenoj leksikografiji. // Studia lexicographica. 2 (3) (2008)

2.3. Primjene korpusne lingvistike

Primijenjena lingvistika opisana je kao uporaba znanja o jeziku za rješavanje problema u stvarnome svijetu. Posljednjih godina lingvisti u području primijenjene lingvistike prihvatili su i priznali prednosti promatranja velikih količina prirodno nastalog jezika u obliku korpusa. Iako korpusi imaju mnoge primijene, a prije svega u forenzičkoj lingvistici, stilistici i kritičkoj analizi diskursa, u nastavku su ukratko navedene i objašnjene dvije češće poticane i nama najbliže primjene korpusne lingvistike: poučavanje jezika i prevođenje.[19]

Poučavanje jezika

Korpusi su utjecali na poučavanje jezika na tri različita načina. Kao prvo, rezultati istraživanja korpusa široko su korišteni za poboljšanje referentnih materijala za učenike, kao što su rječnici i gramatike. Kao drugo, učenike se sve više potiče da istražuju korpuse sami za sebe. Naposljetku, korpusne tehnike primijenjene su kod učenika za učenje dotičnog stranog jezika.[20]

Od sredine 1980-ih korpusi su postali neizostavan resurs za leksikografe i gramatičare. Moderni rječnici za učenike stranih jezika obično više pažnje posvećuju frazeologiji, a posebice kolokacijama, nego su to činili prijašnji. Isto tako, gramatike za učenike više pažnje posvećuju varijacijama registra, govornom jeziku i ulozi leksika u gramatici. U manjoj mjeri promijenili su se i udžbenici koji sada više važnosti daju kolokacijama i frazeologiji. Korpusi su utjecali na metodu, ali i sadržaj poučavanja jezika. Napredne učenike često se potiče da sami pristupe korpusima i „podatkovno upravljanom učenju“, u kojemu se služe korpusom da bi sami došli do vlastitih generalizacija o uporabi jezika.[21]

Naposljetku, sam jezik učenika detaljno se proučava kroz razvoj učeničkih korpusa, odnosno korpusa koji se sastoje od zbirki pisanih ili govorenih tekstova koje su stvorili učenici nekog jezika. To omogućava usporedbu izričaja učenika i izvornih govornika te identifikaciju čestih pogrešaka u jeziku učenika. Uobičajena je metodologija identificirati značajke jezika koje se pojavljuju znatno češće ili rjeđe u učeničkom korpusu nego u sličnim korpusima tekstova izvornih govornika te upotrijebiti takve nesrazmjere kao početnu točku dodatnih kvalitativnih istraživanja. O učeničkim korpusima i njihovoj izradi bit će riječi u nastavku rada.[22]

[19] Hunston, S. Corpus linguistics. // Encyclopedia of Language & Linguistics (Second Edition), 2006.
[20] Ibid.
[21] Ibid.
[22] Ibid.

Prevođenje

Korpusi se mogu upotrebljavati za obučavanje prevoditelja, kao resurs za vježbanje za prevoditelje te kao sredstvo za učenje postupka prevođenja i vrsta izbora s kojima se prevoditelji susreću. Paralelni korpusi uglavnom se upotrebljavaju u tim slučajevima, a postoji i softver koji „poravnava“ dva korpusa tako da se prijevod neke rečenice iz izvornog teksta može jednostavno identificirati. To omogućava proučavanje prijevoda neke riječi u različitim kontekstima. Općenito gledajući, istraživanje paralelnih korpusa naglašava činjenicu da prevoditelji ne prevode riječi nego veće jedinice.[23]

[23] Hunston, S. Corpus linguistics. // Encyclopedia of Language & Linguistics (Second Edition), 2006.

3. Korpusi

3.1. Definicija i razvoj

Prvi projekt širih razmjera koji se bavio prikupljanjem jezika bio je *The Survey of English Usage* Randolpha Quirka započet 1950-ih. Taj je korpus postao temelj gramatike standardnog engleskog jezika narednih desetljeća pod imenom *A Comprehensive Grammar of the English Language*. Kasnije je predstavljao referentu točku za sva empirijska istraživanja jezika, pa tako i za Brown korpus.[24] Brown korpus je milijunski korpus engleskog jezika sastavljen 1960-ih, a poznat je i kao prvi računalno potpomognut jezični korpus.[25] Sastavljen je od uzoraka veličine 2000 riječi iz 500 američkih tekstova iz 15 tekstualnih kategorija. Korpus je pažljivo organiziran, jednostavan za uporabu te je pregledan više puta i stoga gotovo ne sadrži greške. Naknadno je ručno označen prema vrstama riječi. Iako se isprva vjerovalo da će takvi korpusi dati odgovore na mnoga pitanja iz područja gramatike i leksika, lingvisti su ubrzo uvidjeli da korpus s jednim milijunom riječi ne može sadržavati više od malog djelića ukupnog vokabulara jezika.[26] Takvi korpusi danas se nazivaju korpusima uzorcima.[27] Korpusna je lingvistika od tada brzo napredovala, a razvoj i povećanje mogućnosti pohrane na osobnim računalima tijekom 1980-ih i 1990-ih doveli su do izrade velikih višemilijunskih korpusa, često nazvanima korpusima druge generacije.[28] Raspon riječi tih korpusa iznosio je između sedam i tridesetak milijuna pojavnica, [29] a jedni od najboljih primjera tih jezičnih korpusa su Sinclairov *Birmingham Collection* of *English Texts* i BNC (*British National Corpus)*, koji sadrži sto milijuna riječi, a sastavljen je tako da je reprezentativan za engleski jezik i uzor je nacionalnim korpusima drugih jezika.[30] Korpusi treće generacije sadržavaju stotine milijuna riječi i najčešće nastaju kao nusproizvodi suvremenih elektroničkih komunikacijskih sustava. Prema prethodno navedenome možemo zaključiti da je najočitija tendencija u razvoju korpusa rast njihove veličine. Međutim, treba imati na umu da veličina korpusa ne može jamčiti raznovrsnost građe,

[24] Teubert, W.; Čermakova, A. Corpus linguistics : a short introduction. London : New York : Continuum, 2007.
[25] Klobučar Srbić, I. Obol korpusne lingvistike suvremenoj leksikografiji. // Studia lexicographica. 2 (3) (2008)
[26] Teubert, W.; Čermakova, A. Corpus linguistics : a short introduction. London : New York : Continuum, 2007.
[27] Bratanić, M. Korpusna lingvistika na kraju 20. stoljeća i implikacije za suvremenu hrvatsku leksikografiju. // Filologija. 30 – 31 (1998)
[28] Klobučar Srbić, I. Obol korpusne lingvistike suvremenoj leksikografiji. // Studia lexicographica. 2 (3) (2008)
[29] Bratanić, M. Korpusna lingvistika na kraju 20. stoljeća i implikacije za suvremenu hrvatsku leksikografiju. // Filologija. 30 – 31 (1998)
[30] Klobučar Srbić, I. Obol korpusne lingvistike suvremenoj leksikografiji. // Studia lexicographica. 2 (3) (2008)

a vrijednost rezultata dobivenih istraživanjem korpusa izravno je povezana s programskim alatima za pretraživanje i analizu.[31]

Postoji više definicija korpusa, a najopćenitija je ona da je korpus „zbirka tekstova prirodnoga jezika sastavljena po određenim kriterijima“. Tu je jako bitno naglasiti da se i zbirka tekstova sastoji od tekstova skupljenih prema nekim kriterijima, ali i da svaka zbirka tekstova nije korpus.[32] Korpus mora biti reprezentativan, a to znači, u najširem smislu, da se nalazi koje on omogućuje mogu smatrati „općenito valjanima za neki zamisljeni veći korpus ili jezik u cjelini“.[33] Ipak, korpus može biti reprezentativan isključivo za tekstove (jezik) koji se u njemu nalaze, a nikako za jezični univerzum. On ne može predstavljati vokabular općeg jezika jer to nije smislen koncept. Jedina težnja korpusa u smislu vokabulara može biti postizanje zasićenosti. Korpus je zasićen kada stopa rasta vokabulara postane konstantna. Drugim riječima, iako ne postoji trenutak kada će sve riječi biti zastupljene u korpusu, postoji trenutak kada će broj novih riječi na primjerice 1000 dodanih riječi biti stalan.[34] Prema tome, možemo reći da je korpus skup jezičnih odsječaka (s obzirom da ne mora biti sastavljen od cijelih tekstova) koji su „odabrani i skupljeni prema eksplicitnim lingvističkim kriterijima s ciljem da čine jezični uzorak“.[35]

3.2. Vrste korpusa

Korpusi mogu biti različitih veličina, sastavljeni za razne svrhe te se mogu sastojati od tekstova različitih vrsta. Svi su korpusi do određene mjere homogeni; sastavljeni su od tekstova na jednom jeziku ili određenoj varijaciji jezika ili nekom registru itd. Do određene su mjere također i heterogeni s obzirom da su u krajnjem slučaju sastavljeni od većeg broja različitih tekstova. Većina korpusa sadrži uz tekst i informacije poput podataka o samom tekstu, oznaka o vrsti riječi za svaku riječ te informacije sintaksne analize.[36] Jedna od definicija (elektroničkog) korpusa kaže da je to načelna zbirka autentičnih tekstova pohranjena u elektroničkom obliku koja se može upotrijebiti za otkrivanje informacija o jeziku koje se možda ne bi primijetile samom intuicijom. Vrlo važno pitanje jest što potražiti kada se želimo služiti

[31] Bratanić, M. Korpusna lingvistika na kraju 20. stoljeća i implikacije za suvremenu hrvatsku leksikografiju. // Filologija. 30 – 31 (1998)
[32] Klobučar Srbić, I. Obol korpusne lingvistike suvremenoj leksikografiji. // Studia lexicographica. 2 (3) (2008)
[33] Bratanić, M. Korpusna lingvistika na kraju 20. stoljeća i implikacije za suvremenu hrvatsku leksikografiju. // Filologija. 30 – 31 (1998)
[34] Teubert, W.; Čermakova, A. Corpus linguistics : a short introduction. London : New York : Continuum, 2007.
[35] Klobučar Srbić, I. Obol korpusne lingvistike suvremenoj leksikografiji. // Studia lexicographica. 2 (3) (2008)
[36] Hunston, S. Corpus linguistics. // Encyclopedia of Language & Linguistics (Second Edition), 2006.

korpusom? Postoji mnogo vrsta korpusa, ovisno prema kojim kriterijima ih dijelimo.[37] Do samog raslojavanja i specijalizacije korpusa došlo je već devedesetih godina te su se oni tada počeli izrađivati u uže, ciljane svrhe, a standardni ili opći korpusi postali su gotovo nezaobilazni pomoćni alat jezikoslovne struke.[38] Bennett tvrdi da postoji približno osam vrsta korpusa: opći, specijalizirani, učenički, pedagoški, povijesni, paralelni, usporedni i monitor korpus.[39] Sinclair ih pak dijeli na specijalizirane, referentne, monitor, oportunističke, usporedne i paralelne.[40] Koju bi vrstu korpusa trebalo upotrijebiti za neki zadatak ovisi o svrsi korpusa. U nastavku su ukratko objašnjeni i opisani neki od navedenih.

Opći korpusi

Najšira vrsta korpusa je opći korpus. Opći korpusi često su vrlo veliki s više od 10 milijuna riječi i sadrže raznolik jezik pa rezultati dobiveni iz njih mogu biti pomalo općeniti. Iako ne postoji korpus koji bi mogao predstavljati sav mogući jezik, opći korpusi pokušavaju korisnicima dati najširu moguću sliku jezika. Primjeri velikih općih korpusa uključuju *British National Corpus*, *American National Corpus*, Hrvatski nacionalni korpus i *Corpus of Contemporary American English* (COCA). Ti veliki korpusi sadrže pisane tekstove poput novinskih tekstova, članaka iz časopisa, književnih djela i sl. Opći korpusi također mogu sadržavati prijepise govorenog jezika poput neformalnih razgovora, sjednica vlade ili poslovnih sastanka. Opće korpuse potrebno je upotrebljavati ukoliko je cilj doći do općenitih zaključaka o jeziku kao cjelini.[41]

Specijalizirani korpusi

Specijalizirani korpus sadrži tekstove određene vrste, a cilj mu je biti reprezentativan za jezik te vrste. Specijalizirani korpusi mogu biti i veliki i manji, a često su izrađeni s ciljem odgovaranja na vrlo specifična pitanja.[42] S obzirom da su izrađeni za posebnu namjernu, takvi

[37] Bennett R., G. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. University of Michigan Press, 2010.
[38] Bratanić, M. Korpusna lingvistika na kraju 20. stoljeća i implikacije za suvremenu hrvatsku leksikografiju. // Filologija. 30 – 31 (1998)
[39] Bennett R., G. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. University of Michigan Press, 2010.
[40] Bratanić, M. Korpusna lingvistika na kraju 20. stoljeća i implikacije za suvremenu hrvatsku leksikografiju. // Filologija. 30 – 31 (1998)
[41] Bennett R., G. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. University of Michigan Press, 2010.
[42] Ibid.

korpusi ne mogu biti uravnoteženi te se ne mogu služiti u druge svrhe osim one za koju su namijenjeni.[43] Primjeri specijaliziranih korpusa uključuju *Michigan Corpus of Academic Spoken English* (MICASE), koji sadrži isključivo govoreni jezik iz sveučilišnog okruženja, *CHILDES Corpus* koji sadrži jezik koji upotrebljavaju djeca te *Michigan Corpus of Upper Level Student Papers* (MICUSP), koji je zbirka radova iz cijelog niza znanstvenih disciplina. Još je jedan primjer medicinski korpus koji sadrži jezik koji upotrebljavaju medicinske sestre i bolničko osoblje. Navedeni korpusi engleskog jezika često se upotrebljavaju za poučavanje engleskog za posebne namjene (ESP).[44]

Usporedni korpus

Za višejezična istraživanja i primjene potrebni su korpusi na svim jezicima koji slijede iste obrasce sastavljanja te se stoga mogu upotrebljavati za usporedbu jezika. Iz tog razloga takvi korpusi ne mogu biti oportunistički nego se moraju temeljiti na referentnim korpusima. Usporedni korpusi neizostavan su izvor za dvojezične i višejezične leksikone i nove generacije rječnika. Kao značajni primjer može se navesti projekt koji financira Europska komisija, a čiji je cilj izrada usporednih referentnih korpusa (veličine po 50 milijuna riječi) za sve službene jezike Europske unije, uključujući katalonski i irski.[45]

Paralelni korpus

Paralelne korpuse čine tekstovi na jednom jeziku i njihovi prijevodi na druge jezike. Oni služe za pronalaženje prijevodnih ekvivalenata te stoga imaju važnu ulogu u razvoju višejezičnih leksikona. Da bi to bilo moguće, paralelni korpusi moraju se poravnati (ujednačiti) najmanje na razini rečenice, a poželjno bi bilo na razini fraza. Nedostatak paralelnih korpusa očituje se u iskrivljenosti jezika prijevoda, koji ne sadržavaju puni raspon vokabulara i sintakse. Kao rješenje tog problema predlaže se uporaba recipročnog paralelnog korpusa koji sadržava autentične tekstove i prijevode na svim uključenim jezicima. Na taj način omogućila bi se

[43] Bratanić, M. Korpusna lingvistika na kraju 20. stoljeća i implikacije za suvremenu hrvatsku leksikografiju. // Filologija. 30 – 31 (1998)
[44] Bennett R., G. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. University of Michigan Press, 2010.
[45] Teubert, W. Language Resources: The Foundations of a Pan-European Information Society (1995)

dvostruka provjera prijevodnih ekvivalenata. Kolokacija ili frazem brojao bi se kao valjan i prihvatljiv ekvivalent samo u slučaju da se nalazi i u autentičnim tekstovima.[46]

Pedagoški korpus

Pedagoški korpus je korpus koji sadržava jezik koji se upotrebljava u razrednom okruženju. Pedagoški korpusi mogu sadržavati akademske udžbenike, prijepise razgovora u učionici ili bilo koji drugi pisani tekst ili prijepis govorenog jezika s kojim se učenici susreću tijekom obrazovanja. Pedagoški korpusi mogu se upotrebljavati da bi se osiguralo da učenici uče koristan jezik, da bi se istražila dinamika učitelj-učenici ili kao alat za učitelje za samostalnu procjenu napretka.[47]

Oportunistički korpus

Oportunistički korpusi jeftina su alternativa referentnim korpusima. To su zbirke tekstova u elektroničkom obliku koji se mogu prikupiti, pretvoriti i upotrijebiti besplatno ili po vrlo niskoj cijeni. Princip sastavljanja takvih korpusa jest prikupiti sve do čega se može doći i pokušati popuniti „rupe“ čim se prepoznaju. Takvi korpusi u načelu nastaju u okruženjima gdje veličina i pristup korpusu ne predstavljaju problem. Oportunistički su korpusi u načelu virtualni korpusi u smislu da se odabir stvarnog korpusa (iz oportunističkog korpusa) radi prema potrebama određenog projekta.[48]

Monitor korpus

Posebnost monitor korpusa jest u tome što njegova veličina nije konstantna nego se on neprestano nadopunjuje (primjerice godišnje, mjesečno ili čak svakodnevno) novim tekstovima. Na taj način njegova se veličina povećava, pri čemu se pazi da omjer tekstova koji se dodaju u korpus bude stalan. Takvi korpusi u načelu su mnogo veći od korpusa uzoraka. Jedan od najpoznatijih i priznatih monitor korpusa jest *The Bank of English,* koji se od svojeg začetka 1980-ih do danas neprestano povećavao i sada ima više od 500 milijuna riječi. Još je jedan primjer *The Global English Monitor Corpus* čija je izrada započela početkom novog

[46] Teubert, W. Language Resources: The Foundations of a Pan-European Information Society (1995)
[47] Bennett R., G. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. University of Michigan Press, 2010.
[48] Teubert, W. Language Resources: The Foundations of a Pan-European Information Society (1995)

stoljeća u obliku elektroničkog arhiva vodećih novina na svijetu pisanih na engleskom jeziku. Cilj je korpusa pratiti uporabu jezika i semantičke promjene u engleskom jeziku na temelju novina da bi se istražilo dolazi li s vremenom do približavanja ili razilaženja diskursa engleskog jezika u uporabi u Ujedinjenom Kraljevstvu, SAD-u, Južnoj Africi, Australiji i Pakistanu.[49] (67) Možemo reći da je cilj monitor korpusa pratiti razvoj jezika, a njegova posebnost je mogućnost zadržavanja „stanja jezika“ u nekom trenutku za istraživačke svrhe.[50]

[49] McEnery, T.; Xiao, R.; Tono, Y. Corpus-Based Language Studies: an advanced resource book. New York : Routledge, 2006.
[50] Sinclair, J. Corpus, concordance, collocation. Oxford : Oxford University Press, 1991., str. 25-26

4. Učenički korpusi

4.1. Definicija i uporaba

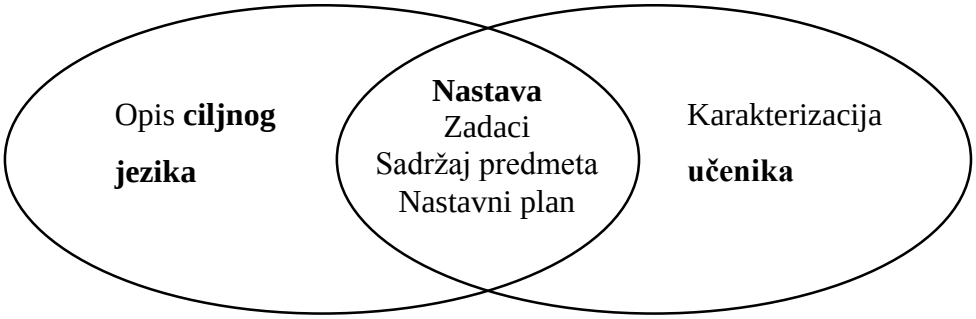
Učenički korpusi elektroničke su zbirke tekstova koje su proizveli učenici jezika kojima dotični jezik nije materinji. Jedna od glavnih uporaba takvih korpusa jest pružanje dokaza o učenju jezika i usporedba načina na koji se jezikom služe neizvorni i izvorni govornici ili govornici s drugim materinskim jezikom.[51] Za učenike jezika jedna od opće prihvaćenih definicija kaže da su govornici koji uče jezik koji nije njihov prvi jezik, ali niti jezik koji je dodatni službeni jezik u njihovoj zemlji prebivališta. Tom uskom definicijom zapravo su obuhvaćeni isključivo učenici „stranog jezika“, bez učenika „drugog“ i „prvog jezika“. S obzirom da se istraživanja koja se služe učeničkim korpusima općenito bave „međujezikom“, odnosno „prijelaznim“ jezikom i čimbenicima koji na njega utječu, istraživači su s vremenom počeli pojam *učenik jezika* upotrebljavati i za učenike drugog i materinskog jezika.[52]

Učenički korpusi osobito su povezani s područjem usvajanja drugog jezika (SLA – *second language acquisition*) koje pokušava dokučiti mehanizme usvajanja stranog/drugog jezika, te područjem poučavanja stranog jezika (FLA – *foreign language teaching*) čiji je cilj poboljšati učenje i poučavanje stranih/drugih jezika.[53] Glavna tema SLA-a već je spomenuti međujezik. Radi se o ideji da jezik koji učenici upotrebljavaju nije samo rezultat razlika između jezika koje već znaju i jezika koji uče, nego da se radi o potpunom samostalnom sustavu koji ima vlastita pravila. Prema toj teoriji, međujezik se postepeno razvija kako učenik dolazi sve više u doticaj s jezikom koji se uči.[54]

Mark ističe da su određeni čimbenici koji imaju ulogu u učenju i poučavanju jezika dobili kroz povijest više pozornosti od drugih.[55] Najpopularniji pristupi poučavanju jezika obično se bave s tri dijela prikazana na slici 1. Uloženi su veliki napori ne bi li se bolje opisao ciljni jezik, a povećan je i interes za varijable povezane s učenikom, kao što su motivacija, načini učenja,

[51] Mikelić Preradović, N.; Berać, M.; Boras, D. Learner Corpus of Croatian as a Second and Foreign Language, 2015.
[52] Glaznieks, A.; Nicolas, L.; Stemle, E.; Abel, A.; Lyding, V. Establishing a Standardised Procedure for Building Learner Corpora, 2014.
[53] Granger, S. A Bird's-eye view of learner corpus research, 2002.
[54] Second-language acquisition. // Wikipedia: the free encyclopedia. (19. 7. 2017.)
[55] Mark, K.L. The Significance of Learner Corpus Data in Relation to the Problems of Language Teaching, 1998.

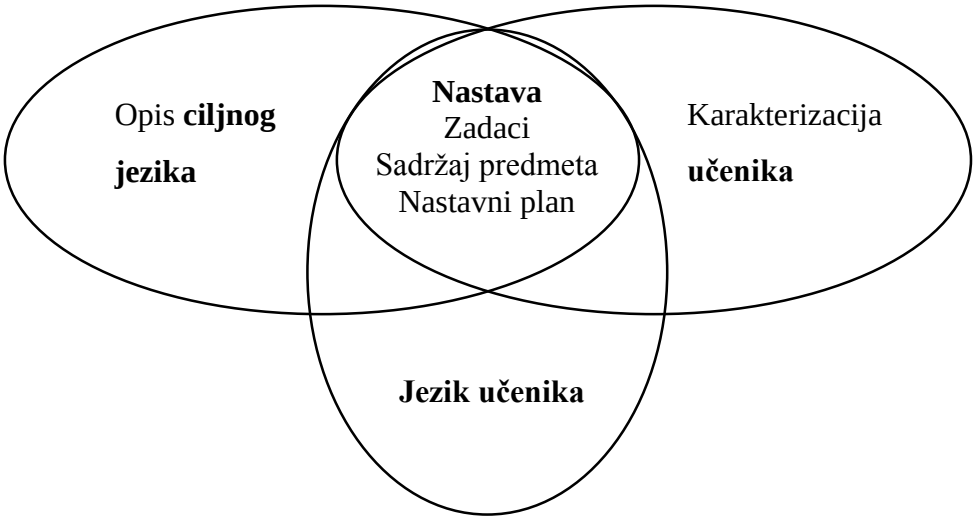
potrebe, stav itd. Bolje razumijevanje ciljnog jezika i učenika dovelo je do razvoja učinkovitijih zadataka, sadržaja predmeta i nastavnih planova u području poučavanja jezika.[56]



Slika 1.: glavni problemi u poučavanju jezika[57]

Ono što vidljivo nedostaje jest rezultat ili učinak učenika. Iz prikazanoga je jasno da jezik učenika nije zauzimao ključnu ulogu. Mark dodaje kako je u potpunoj suprotnosti sa zdravim razumom temeljenje nastave na ograničenim podacima o učenicima, pritom ignorirajući u svim aspektima pedagogije, na razini od zadatka do nastavnog plana, poznavanje jezika učenika.[58] Stoga ne čudi postepeni okret i prelazak istraživača u područjima usvajanja drugog jezika i poučavanja stranog jezika prema učeničkim korpusima i vrstama opisa koje oni mogu ponuditi. Upravo učenički korpusi mogu istraživačima ponuditi rezultate učenja učenika u obliku velikih količina podataka u elektroničkom obliku dobivenih u kontroliranim uvjetima, koji se mogu obraditi i analizirati na više razina s pomoću softverskih lingvističkih alata. Na slici 2. prikazano je kako bi uključivanje rezultata učenika dovelo do boljeg razumijevanje tri područja iz slike 1.[59]

[56] Granger, S. A Bird's-eye view of learner corpus research, 2002.
[57] Mark, K.L. The Significance of Learner Corpus Data in Relation to the Problems of Language Teaching, 1998.
[58] Ibid.
[59] Granger, S. A Bird's-eye view of learner corpus research, 2002.



Slika 2.: usmjeravanje na produkciju učenika[60]

Iako su korpusi samo jedan od izvora dokaza o uporabi jezika, njihova uloga u istraživanju značajki poput frekvencije ne može se zanemariti ili nadomjestiti. Frekvencija je aspekt jezika kojega smo intuitivno svjesni u vrlo skromnoj mjeri, ali upravo ona ima važnu ulogu u mnogo lingvističkih primjena koje zahtijevaju znanje o tome što je u jeziku moguće, ali i što ima veću vjerojatnost pojavljivanja. Velika i očita prednost računalnih korpusa svakako je u mogućnosti provođenja kvantitativnih analiza. Provođenje kvantitativnih usporedbi širokog raspona lingvističkih značajki u korpusima koji predstavljaju različite varijacije jezika moguće je otkriti kako se različite značajke pojavljuju zajedno u različitim obrascima pojavljivanja. Takva istraživanja omogućila su mnogo bolje opise razlika među registrima (formalni i neformalni jezik, akademski tekstovi, novine itd.) i dijalektima jezika (primjerice britanski i američki engleski, jezik muškaraca i žena).[61]

4.2. Razvoj i pregled učeničkih korpusa u svijetu

Istraživanje neizvornih varijanti jezika relativno je nov smjer u području korpusne lingvistike: istraživači i izdavači počeli su prikupljati korpuse engleskog jezika (kao najveći svjetski jezik predvodnik je u istraživanjima) neizvornih govornika tek krajem 1980-ih i početkom 1990-ih.[62] Jedan od prvih i glavnih korpusa je *International Corpus of Learner English* (ICLE) koji je

[60] Mark, K.L. The Significance of Learner Corpus Data in Relation to the Problems of Language Teaching, 1998.
[61] Granger, S. A Bird's-eye view of learner corpus research, 2002.
[62] Ibid.

razvijen na sveučilištu Université catholique de Louvain u Belgiji.[63] Prva verzija sastojala se od 2,5 milijuna riječi u obliku raspravljačkih eseja koje su napisali učenici engleskog jezika iz niza europskih zemalja s različitim materinskim jezicima poput španjolskog, talijanskog, francuskog, ruskog itd. Korpus je organiziran u manje podkorpuse prema materinskim jezicima učenika, što omogućava kontrastivnu analizu međujezika između različitih skupina učenika, ali i usporedbu s jezikom izvornih govornika s pomoću ekvivalentnog korpusa engleskog jezika izvornih govornika *Louvain Corpus of Native English Essays* (LOCNESS), koji je sastavljen od eseja učenika iz UK-a i SAD-a. Kasnije je objavljena proširena verzija s 3,7 milijuna riječi sa sveukupno 16 materinskih jezika, koji su tada uključivali i jezike poput japanskog, kineskog, turskog i dr. Objava ICLE-a može se smatrati početnom točkom šire uporabe učeničkih korpusa te je sam korpus potaknuo mnoga istraživanja u području učeničkih korpusa.[64] Vidljivo je to iz oko 400 radova izrađenih na temelju ICLE-a[65] i više od 100 učeničkih korpusa pisanog i govorenog jezika diljem svijeta (popis dostupan na https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html). Međutim, iako su jezici poput njemačkog, francuskog, španjolskog, arapskog i sl. prepoznali rastuću važnost takvih korpusa, veliki učenički korpusi (s više od milijun riječi) rijetki su za „manje“ jezike.[66]

Jedan od primjera učeničkih korpusa izrađenih prema principima ICLE-a jest *Written Corpus of Learner English* (WriCLE) koji je sastavljen na *Universidad Autónoma de Madrid*. Radi se o korpusu španjolskih studenata prve i treće godine preddiplomskog studija engleskog jezika. Za razliku od ICLE-a, razina znanja jezika utvrđena je prema Zajedničkom europskom referentnom okviru za jezike (ZEROJ), što je veoma bitno za vrijednost analiza provedenih na temelju tekstova iz korpusa.[67] Korpus se trenutačno sastoji od 750.000 riječi i besplatno je dostupan svima, zajedno sa sučeljem za pretraživanje i dohvaćanje rečenica.[68] Za španjolski je još bitnije navesti *Corpus Escrito del Español como L2* (CEDEL2), sastavljen od tekstova učenika španjolskog jezika i podkorpusa izvornih govornika za svrhe usporedbe.[69] Sadrži više od 700.000 riječi od otprilike 2300 sudionika (oko 1700 čine izvorni govornici engleskog jezika

[63] Mikelić Preradović, N.; Berać, M.; Boras, D. Learner Corpus of Croatian as a Second and Foreign Language, 2015.
[64] Lozano, C.; Mendikoetxea, A. Learner corpora and Second Language Acquisition: The design and collection of CEDEL2, 2013.
[65] Ibid.
[66] Mikelić Preradović, N.; Berać, M.; Boras, D. Learner Corpus of Croatian as a Second and Foreign Language, 2015.
[67] Lozano, C.; Mendikoetxea, A. Learner corpora and Second Language Acquisition: The design and collection of CEDEL2, 2013.
[68] Written Corpus of Learner English. // Dostupno na: http://web.uam.es/proyectosinv/woslac/Wricle/ (19.7.2017.)
[69] Ibid.

koji uče španjolski, dok oko 600 čine izvorni španjolski govornici). Tekstovi su prikupljeni putem internetskog obrasca koji sudionici moraju ispuniti.[70]

Danas već postoji prevelik broj učeničkih korpusa u raznim oblicima da bi ih sve ovdje naveo, ali spomenut ću samo neke od njih radi boljeg uvida u razvoj situacije tijekom zadnjih 15 – 20 godina. Za njemački jezik tako postoje pisani korpusi *The Corpus of Learner German* (CLEG13) s više od 300.000 riječi, *The Telecollaborative Learner Corpus of English and German Telekorp* s 1,5 milijuna riječi te korpus govornog jezika *The Learning the Prosody of a Foreign Language* (LeaP) sastavljen od govornoga jezika 62 osobe. Za francuski jezik možemo među ostalima istaknuti *The Learner Corpus French* (LCF) s 500.000 riječi od izvornih govornika nizozemskog jezika, *French Interlanguage Database* (FRIDA) te govorne korpuse *French Learner Language Oral Corpora* (FLLOC) i *Interphonologie du Français Contemporain* (IPFC). Za talijanski su jezik dostupni govorni korpus *The Lexicon of Spoken Italian by Foreigners* (LIPS) sa 700.000 riječi te pisani korpus *Varietà di Apprendimento della Lingua Italiana: Corpus Online* (VALICO) s 570.000 riječi.[71] Od azijskih jezika možemo istaknuti pisani korpus kineskog jezika *The Jinan Chinese Learner Corpus* (JCLC) sastavljen od tekstova izvornih govornika 50 različitih materinskih jezika, veličine 6 milijuna kineskih znakova i 9000 tekstova.[72]

Što se tiče jezika bliskih hrvatskom jeziku, odnosno slavenskih jezika koji imaju bogatu morfologiju, situacija je znatno lošija te možemo istaknuti svega nekoliko korpusa. Najveći i vjerojatno najpoznatiji od njih je CzeSL, korpus češkog kao drugog/stranog jezika, koji je prvi korpus izrađen za neki inflektivni jezik. Planirana veličina korpusa je dva milijuna riječi, a uključeni su učenici sa svim razinama znanja češkog jezika (prema Zajedničkom europskom referentnom okviru za jezike). Sastoji se od četiri podkorpusa prema materinskom jeziku učenika: ruski podkorpus predstavlja međujezik učenika sa slavenskim materinskim jezikom, vijetnamski podkorpus predstavlja brojne manje skupine učenika s materinskim jezikom koji nije blizak češkom, tu je također i podkorpus romske manjine te podkorpus sa svim ostalim materinskim jezicima. Svaki podkorpus dalje je podijeljen na pisani i govorni.[73] PiKUST je učenički korpus slovenskog jezika veličine 35.000 riječi iz 128 tekstova koje je napisalo 119 učenika s 18 različitih materinskih jezika. Korpus je prikupljen oportunistički te je većina jezika

[70] Corpus Escrito de Espanol L2. // Dostupno na: http://www.uam.es/proyectosinv/woslac/collaborating.htm (19. 7. 2017.)
[71] Learner corpora around the world. // Université catholique de Louvain. Dostupno na: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (19. 7. 2017.)
[72] Ibid.
[73] Hana, J.; Rosen, A.; Škodova, S.; Štindlova, B. Error-tagged Learner Corpus of Czech, 2010.

zastupljena s manje od 2000 riječi, što znači da su više ilustrativni nego temelj za istraživanja. Više od polovice (67%) korpusa čine tekstovi izvornih govornika hrvatskog i srpskog jezika.[74] Za ruski jezik postoji *Russian Learner Corpus of Academic Writing* (RULEC), longitudinalni korpus učenika ruskog jezika od 750.000 riječi. Sastavljen je od približno 3800 tekstova (veličine od jednog paragrafa do istraživačkih radova od 8 stranica) 36 američkih učenika kojima je ruski strani ili nasljedni jezik.[75]

Za hrvatski je jezik važno spomenuti Učenički korpus hrvatskog kao drugog i stranog jezika koji je sastavljen od dva podkorpusa (pisani i govorni) koji sadržavaju tekstove, odnosno govor učenika na svim razinama znanja jezika prema Zajedničkom europskom referentnom okviru za jezike (od A1 do C2). Pisani podkorpus naziva se *CROatian Learner TExt Corpus* (CROLTEC) te sadržava 1.054, 287 riječi. Velika većina tekstova digitalizirana je skeniranjem i prepisivanjem, dok je znatno manji dio izvorno digitalnog oblika. Tijekom prepisivanja zadržani su ispravci koje su učenici unosili sami. Osim toga, uz svaki tekst uključeni su sociolingvistički metapodaci poput dobi, spola, nacionalnosti, razine znanja, materinskog jezika i sl. Dokumenti su pretvoreni u TEI XML format, a korpus se nalazi na TEITOK sustavu (http://teitok.iltec.pt/croltec/index.php?action=home) te će nakon dovršetka MSD označavanja biti dostupan za pretraživanje. Također je razvijena i shema za označavanje pogrešaka te će i ta obrada biti naknadno dodana. Korpus bi trebao omogućiti dubinsku analizu učeničkog jezika i opis odstupanja od standardnog hrvatskoga jezika. Govorni podkorpus *CROatian Learner Speech Corpus* (CROLSEC) sadržava zvučne zapise pročitanog teksta te zapise spontanog govora. Prikupljeno je 9 sati i 25 minuta zvučnih zapisa od 144 različita učenika. Analizom tog korpusa trebalo bi doći do važnih otkrića u području pogrešaka i propusta koje u govoru čine učenici hrvatskog kao drugog ili stranog jezika. Posebna pozornost posvetit će se fonetskom opisu međujezika učenika sa materinskim jezikom iz slavenske skupine jezika. U planu je prepisivanje zvučnih zapisa u tekst i spajanje s pisanim podkorpusom.[76] Za kraj ću spomenuti Korpus srpskog kao stranog jezika čija je izrada započela 2014. godine. Sastoji se prvenstveno od učeničkih sastavaka pisanih u sklopu ispita i nastavnih aktivnosti u Centru za srpski jezik kao strani jezik na Filološkom fakultetu u Beogradu. Iz praktičnih razloga kriteriji za uključivanje tekstova definirani su što je moguće šire. Najveća je poteškoća u izradi nedostatak

[74] Stritar, M. Slovene as a Foreign Language : The Pilot Learner Corpus Perspective, 2009.
[75] Mikelić Preradović, N.; Berać, M.; Boras, D. Learner Corpus of Croatian as a Second and Foreign Language, 2015.
[76] Ibid.

građe – problem s kojim se suočavaju svi ili gotovo svi „manji“ jezici, koji nemaju dovoljno velik broj učenika, osobito na višim razinama učenja.[77]

[77] Miličević, M. Korpus srpskog kao stranog jezika (KSKS): Opis građe i plan izrade, 2016.

5. Izrada učeničkog korpusa

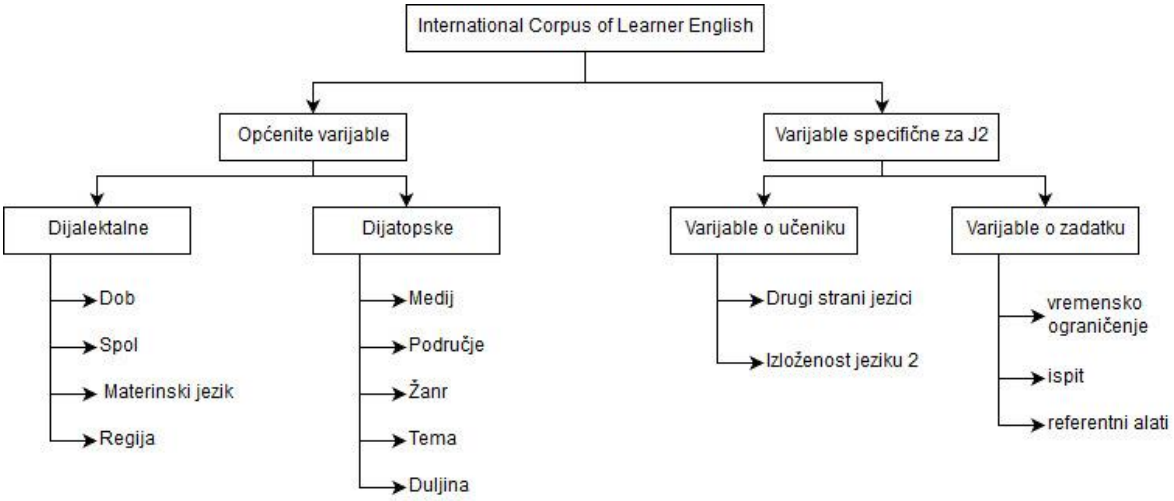
Jedna od prednosti korpusnog istraživanja jezika učenika jest i mogućnost dijeljenja tih korpusa s drugim istraživačima koji mogu detaljno provjeriti sve eventualno dobivene rezultate. Tono upozorava da upravo u tom pogledu mnogi istraživači griješe prilikom izrade prvog učeničkog korpusa jer prije početka izrade ne posvete dovoljno pažnje nekim pitanjima povezanim s projektiranjem. Prikupljanje podataka na oportunistički način bez dovoljne razine kontrole i dokumentacije o raznim varijablama neće dati korpus koji je vrlo upotrebljiv (u smislu lingvističkih analiza).[78]

Izradu učeničkog korpusa možemo podijeliti u tri koraka: projektiranje, prikupljanje podataka i analizu prikupljenih podataka. U koraku projektiranja određuju se varijable korpusa. Primjerice, pažnja može biti usmjerena na varijable povezane s jezikom, zadatkom i/ili učenikom. U koraku prikupljanja podataka prikupljaju se sirovi tekst, zvuk te informacije koje će se označiti uz tekst, kao što su podaci o učenicima i pogreškama. U koraku analiziranja prikupljenih podataka obavljaju se osnovne analize poput deskriptivnih statističkih analiza kvalitativnih analiza kako bi se potvrdila valjanost prikupljenih podataka.[79]

5.1. Varijable u projektiranju

Kriteriji projektiranja vrlo su bitni za jezik učenika. Nasumična zbirka heterogenih učeničkih tekstova ne čini učenički korpus. Oni se moraju sastaviti u skladu sa strogim kriterijima projektiranja, pri čemu su neki isti kao za korpuse izvornih govornika, dok su neki specifični za učeničke korpuse i odnose se primjerice na učenika i zadatak. Korisnost učeničkog korpusa izravno je proporcionalna pažnji koja je posvećena kontroliranju i kodiranju varijabli.[80]

[78] Tono, Y. Learner corpora: design, development and applications. 2003.
[79] Kotani, K.; Yoshimi, T.; Nanjo, H.; Isahara, H. Corpus Materials for Constructing Learner Corpus Compiling Speaking, Writing, Listening, and Reading Data, 2012.
[80] Granger, S. A Bird's-eye view of learner corpus research, 2002.



Slika 3.: varijable primijenjene u izradi ICLE-a[81]

S obzirom da istraživače zanimaju različiti aspekti učeničkog jezika, logično je da će se dizajn učeničkog korpusa razlikovati od projekta do projekta, iako se ne smije zanemariti značajan utjecaj već spomenutog ICLE-a. U tablici 1. prikazana su neka pitanja koja treba uzeti u obzir prilikom izrade učeničkog korpusa. Podijeljena su u tri glavne kategorije: 1) kriteriji povezani s jezikom (npr. oblik, medij, žanr, tema); 2) kriteriji povezani sa zadatkom (npr. longitudinalni ili poprečni, spontani ili pripremljeni); 3) kriteriji povezani s učenikom (npr. drugi ili strani jezik, dob, spol, materinski jezik i sl.).[82]

Tablica 1.: razmatranja o dizajnu prije izrade učeničkog korpusa[83]

Vrste značajki		
Jezik	**Zadatak**	**Učenik**
Oblik [pisani/govorni]	Prikupljanje podataka [poprečno/longitudinalno]	Unutarnje kognitivne [dob/kognitivni stil]
Žanr [pismo/dnevnik/fantastika/esej]	Elicitacija [spontano/pripremljeno]	Unutarnje afektivne [motivacija/stav]
Stil [pripovijedanje/raspravljanje]	Uporaba referenci [rječnik/izvorni tekst]	Materinski jezik (J1)
Tema [općenito/slobodno vrijeme/itd.]	Vremensko ograničenje [fiksno/slobodno/domaća zadaća]	Okolnosti za J2 [drugi/strani]/[razred/godina]
		Poznavanje J2 [rezultat na standardiziranom ispitu]

[81] Granger, S. Computer learner corpus research : current status and future prospects, 2004.
[82] Tono, Y. Learner corpora: design, development and applications. 2003.
[83] Ibid.

Osim toga, učenički korpus može se izraditi u nizu formata. Tako može biti sirovi korpus, odnosno korpus s običnim tekstom bez dodatnih dodanih značajki, ili označeni korpus tj. korpus obogaćen lingvističkim ili tekstualnim informacijama kao što su gramatičke kategorije ili sintaktičke strukture. Označeni korpus trebao bi se u idealnom slučaju temeljiti na standardiziranom softveru za označavanje da bi se omogućila usporedivost označenog učeničkog korpusa s označenim korpusom izvornih govornika.[84]

Dodatne odluke u projektiranju učeničkog korpusa povezane su s tipologijom. Tipologija se često definira u okviru dihotomija, od kojih su četiri osobito važne za učeničke korpuse. Učenički korpusi prvenstveno su **jednojezični**, iako već postoji određeni broj korpusa s uzorkom J1 i jezikom izvornih govornika. Druga dihotomija odnosi se na jezik, pri čemu treba istaknuti da većina učeničkih korpusa sadržava **općeniti jezik**, dok manji broj korpusa sadržava specijalizirani jezik. Osim toga, učenički korpusi uglavnom su **sinkronijski,** odnosno opisuju učenički jezik u određenom trenutku. Postoji vrlo malo longitudinalnih korpusa, tj. korpusa koji obuhvaćaju razvoj učeničkog jezika. Razlog za to vrlo je jednostavan – za takve korpuse bilo bi potrebno učenike pratiti mjesecima ili godinama. Obično istraživači koji su zainteresirani za razvoj učeničkog poznavanja jezika prikupljaju „kvazi-longitudinalne“ podatke, odnosno prikupljaju podatke od homogenih skupina učenika na različitim razinama učenja (npr. usporedba znanja studenata prve i treće godine). Odabir između pisanog i govornog jezika predstavlja posljednju dihotomiju. Problemi povezani s prikupljanjem podataka i sastavljanjem učeničkih korpusa posebno su vidljivi u slučaju prikupljanja govornog jezika te zbog toga postoji mnogo više korpusa **pisanog jezika.**[85] Iz tog razloga postupak u sljedećem poglavlju o prikupljanju i obradi podataka odnosit će se prvenstveno na pisani jezik.

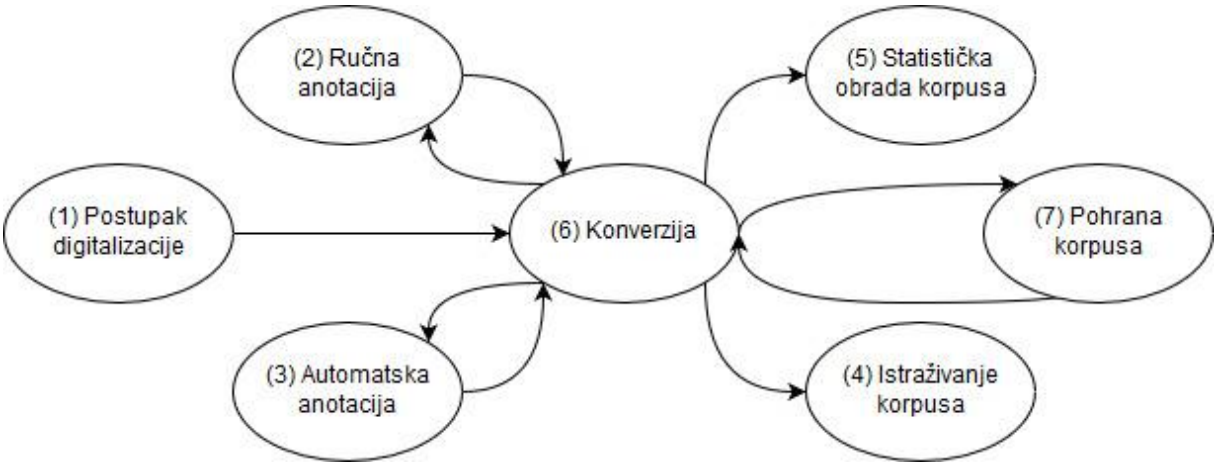
5.2. Prikupljanje i obrada podataka

Da bi neki tekst mogao biti dio učeničkog korpusa, taj uzorak jezika mora se sastojati od neprekinutog dijela diskursa, a ne izoliranih rečenica ili riječi. Iz tog razloga ne može postojati „korpus pogrešaka“. Dakle učenički korpus sastoji se od neprekinutih dijelova diskursa koji sadrže ispravnu i neispravnu upotrebu jezika.[86] Potencijalni je problem odabir korpusnih materijala koji ispravno prikazuju učenikovo znanje stranog jezika. Iz tog razloga u izradi

[84] Granger, S. A Bird's-eye view of learner corpus research, 2002.
[85] Ibid.
[86] Ibid.

učeničkih korpusa obično se upotrebljavaju materijali dobiveni iz lingvističkih vježbi kao što su eseji i jezični ispiti.[87]

Nakon određivanja i dokumentacije svih mogućih varijabli (u idealnom slučaju) u prethodnom koraku, odnosno fazi projektiranja i dizajna učeničkog korpusa, dolazi faza samog prikupljanja potrebnih uzoraka jezika (tekst, govor) te obrada tih podataka. S obzirom da se radi o poprilično opširnom postupku, drugi korak ili fazu možemo podijeliti u manje dijelove. Glaznieks i ostali tako predlažu postupak prikazan na slici 4.



Slika 4.: apstraktni postupak izrade korpusa[88]

Osim toga, u svom radu navode i nekoliko ključnih zahtjeva i objašnjavaju kako se stavke iz navedenog postupka izrade korpusa odnose na te zahtjeve te kako se u njihovu svrhu zapravo primjenjuju. Zahtjevi su tako podijeljeni u dvije skupine: zahtjeve povezane s korpusom (proširive anotacije korpusa, anotacije korpusa visoke kvalitete, pretraživost korpusa) te zahtjeve povezane s postupkom izrade korpusa (učinkovit postupak za ljudski rad, dinamički ocjenjiv i prilagodljiv postupak te formaliziran i ponovljiv postupak).[89]

Proširive anotacije korpusa podrazumijevaju da se tijekom različitih faza obrade dodaju različite razine anotacije. U početku se korpus može označiti informacijama o vizualnom izgledu dokumenata, kao što su grafički raspored (zaglavlje, paragrafi itd.) i samostalne ispravke (umetanje, brisanje). Zatim je moguće dodati informacije o odstupanju od standardnog

[87] Kotani, K.; Yoshimi, T.; Nanjo, H.; Isahara, H. Corpus Materials for Constructing Learner Corpus Compiling Speaking, Writing, Listening, and Reading Data, 2012.
[88] Glaznieks, A.; Nicolas, L.; Stemle, E.; Abel, A.; Lyding, V. Establishing a Standardised Procedure for Building Learner Corpora., 2014.
[89] Ibid.

pisanog jezika (npr. ortografske i morfosintaktičke pogreške). Da bi se omogućila proširivost anotacija u korpusu, razine anotacije moraju se u različitim fazama obrade (komponente ručne (2) i automatske (3) anotacije te prije njih digitalizacije (1)) dodavati na strukturiran i sustavan način. Osim toga, komponente konverzije (6) i pohrane podataka (7) moraju moći prihvatiti sve podatke iz prethodne tri komponente.[90]

Anotacije korpusa visoke kvalitete predstavljaju temelj za precizne analize učeničkog jezika u učeničkim korpusima. To je bitno već za same razine anotacije jer male razlike mogu predstavljati razliku između učenika ili skupina učenika.. Dodavanjem novih razina anotacije kvaliteta postaje još bitnija jer se pogreške mogu brzo proširiti i lažno povećati razlike između pojedinaca i skupina. Temelj anotacija visoke kvalitete nalazi se u precizno definiranom protoku podataka među komponentama u postupku te mogućnosti ponavljanja koraka obrade podataka radi poboljšanja kvalitete.[91]

Za istraživanje učeničkog korpusa lingvisti moraju biti u mogućnosti provoditi praktična **pretraživanja i analize korpusa**. Kod većih korpusa provođenje neautomatiziranih analiza zahtijeva previše vremena i rada te je podložno pogreškama. Osim toga, lingvisti moraju biti u mogućnosti ispitati svoje ideje i izraditi statistike na jednostavan i dinamičan način. Iz tog razloga važno je da se korpus može pretraživati s pomoću sofisticiranih upita te da se statistike mogu izraditi na dobivenim rezultatima, uzimajući u obzir različite razine anotacije. Navedene mogućnosti obuhvaćene su komponentama za istraživanje korpusa (4) i statističke obrade korpusa (5). Osim toga, komponenta konverzije (6) osigurava interoperabilnost između te dvije komponente te podataka iz postupka digitalizacije (1) i komponenti anotacije (2 i 3).[92]

Glaznieks i ostali ističu da je među resursima potrebnim za izradu učeničkog korpusa **ljudski rad** najbitniji i najčešće najrjeđi. U skladu s time, svaki način poboljšavanja ručnog rada ili izbjegavanja ponavljanja, odnosno automatizacija ručnog rada, predstavlja značajan učinak na veličinu konačnog korpusa i njegove razine anotacije, a prema tome i valjanost argumenata koji se iz njega dobiju. Učinkovitost postupka za ljudski, odnosno ručni rad odnosi se prvenstveno na komponente ručne anotacije (2) i postupka digitalizacije (1), ukoliko uključuje ručni rad. Radni postupak trebao bi moći brzo otkrivati sve pogreške u anotaciji. U tu svrhu komponente (4) i (5) trebaju omogućiti korisnicima pretraživanje korpusa radi pogrešaka u prijepisu i

[90] Glaznieks, A.; Nicolas, L.; Stemle, E.; Abel, A.; Lyding, V. Establishing a Standardised Procedure for Building Learner Corpora., 2014.
[91] Ibid.
[92] Ibid.

anotaciji te implementaciju metoda i testova za otkrivanje takvih pogrešaka. Polu-automatska anotacija učinkovito objedinjuje računalne (3) i ljudske resurse (1, 2) te smanjuje ljudski rad. Osim toga, opcionalni sustav za praćenje verzija može pomoći u vraćanju prethodnih verzija anotacija i brz nastavak rada.[93]

Prilikom izrade korpusa korisno je pratiti kvalitetu i kvantitetu prijepisa i anotacija u svrhu ranog prepoznavanja problema. Primjerice ako je ujednačenost među označivačima na nekoj razini anotacija niska. S obzirom da je takve probleme teško unaprijed predvidjeti, postupak bi trebao olakšati redovitu ocjenu korpusa i po potrebi ispravak i proširenje svih vrsta anotacija. To predstavlja **dinamički ocjenjiv i prilagodljiv postupak**. U tu svrhu upotrebljavaju se komponente za istraživanje (4) i statističku obradu korpusa (5) pomoću kojih se provode kvalitativne i kvantitativne analize. U slučaju da rezultati pokažu neusklađenost s unaprijed zadanim kriterijima, anotacije je moguće prilagoditi s pomoću komponenti za ručnu i automatsku anotaciju (2, 3). Komponente konverzije i pohrane (6, 7) također pomažu u dinamičkom postupku izmjene.[94]

Formaliziran i ponovljiv postupak označava činjenicu da bi postupak trebao dati nacrt za izradu korpusa, odnosno postupak treba biti formaliziran na takav način da su (glavne) odluke istaknute, tako da identični ciljevi i odluke u projektiranju korpusa daju identične rezultate. Osim toga, drugi istraživači morali bi moći ponoviti rezultate postupka izrade korpusa pod uvjetom da su upoznati s ciljevima i odlukama u projektiranju te imaju pristup posrednim podacima.[95]

Što se tiče same anotacije i obrade jezika, osim značajki navedenih u prethodnom poglavlju, podaci iz učeničkog korpusa mogu se podvrgnuti obradi s pomoću tehnika obrade prirodnog jezika. Dostupne vrste takve obrade navedene su u tablici 2.

[93] Glaznieks, A.; Nicolas, L.; Stemle, E.; Abel, A.; Lyding, V. Establishing a Standardised Procedure for Building Learner Corpora., 2014.
[94] Ibid.
[95] Ibid.

Tablica 2.: obrada učeničkih podataka[96]

Ekstra-tekstualni podaci	Podaci u zaglavlju (učenik/jezik/varijable zadatka)
Razina transkripcije	Ortografski (+fonemski/fonetski za govorne korpuse)
Razine anotacije	Određivanje granica rečenica Tokenizacija Označavanje vrsta riječi Parsiranje (banke stabala) Semantičko označavanje (značenje riječi/semantički odnosi i kategorije) Označavanje diskursa Označavanje pogrešaka Prozodijsko označavanje Anaforičko označavanje

U zajednici korpusne lingvistike sve više raste i svijest o potrebnosti standardizacije formatiranja i anotacije korpusa koliko god je to moguće. Primjer toga je i razvoj standarda kao što su primjerice TEI, CES/XCES, EAGLES, ATLAS, TUSNELDA i MATE te će u nastavku biti ukratko objašnjeni neki od njih.

5.2.1. CES

Corpus Encoding Standard (CES) nastao je zbog potrebe za skupom standarda za kodiranje korpusa uslijed velikog porasta broja projekata koji su se bavili prikupljanjem jezika, odnosno izradom korpusa. Razvili su ga zajednički Expert Advisory Group on Language Engineering Standards (EAGLES), MULTEXT te Vassar/CNRS. Ideja je bila stvoriti standard za kodiranje korpusa koji je optimalno prilagođen za uporabu u području jezičnog inženjeringa, a koji bi služio kao opće prihvaćeni skup standarda kodiranja za poslove temeljene na korpusima. Općeniti cilj bio je identifikacija minimalne razine kodiranja koju korpusi moraju ostvariti da bi se smatrali standardiziranima u smislu opisnog prikaza (označavanje strukturalnih i lingvističkih informacija). CES također pruža konvencije kodiranja za još šire kodiranje i lingvističko označavanje, kao i općenitu arhitekturu (da bi u najvećoj mogućoj mjeri bio prikladan za uporabu u tekstualnim bazama podataka) radi prikaza korpusa označenih

[96] Tono, Y. Learner corpora: design, development and applications. 2003.

lingvističkim značajkama. Njegova je namjena uporaba u kodiranju korpusa koji se koriste kao resursi u jezičnom inženjeringu, uključujući sva područja obrade prirodnog jezika, strojnog prevođenja, leksikografije itd. Sam CES baziran je na SGML-u i sukladan je sa specifikacijama TEI smjernica za kodiranje i razmjenu elektroničkog teksta (*TEI Guidelines for Electronic Text Encoding and Interchange*) koji je razvila Inicijativa za kodiranje teksta (*Text Encoding Initiative*). U smislu standardizacije mogu se razlikovati tri razine.[97]

Razina metajezika – standardizacija na razini metajezika određuje oblik sintaksnih pravila i osnovne mehanizme shema označavanja. Međutim, na toj razini nije definirano samo označavanje (nazivi oznaka, dopuštene sekvence oznaka itd.). Koristeći se DTD (*Document Type Definition*) dokumentom korisnik može definirati nazive oznaka, „modele dokumenata" koji određuju odnose među oznakama i „sintaktička" pravila za uporabu oznaka. To predstavlja **sintaktičku razinu**. Posljednja je razina **semantička razina**. Semantika u označavanju obično nije formalna, odnosno oslanja se na korisnika da upotrijebi neku oznaku na odgovarajući način. CES pokušava standardizirati semantičku razinu za elemente koji su najvažniji u primjenama jezičnog inženjeringa, osobito lingvističkih elemenata.[98]

```
<cesDoc ...>
<cesHeader ...>
...
<wordCount>828388</wordCount>
<byteCount units="bytes">5418715</byteCount>
...
<langUsage>
    <language iso639="afr" id="af">Afrikaans</language>
</langUsage>
</cesHeader>
<text>
<body id="Bible" lang="af">
<div id="b.GEN" type="book">
<div id="b.GEN.1" type="chapter">
    <seg id="b.GEN.1.1" type="verse">
        In die begin het God die hemel en die aarde geskape.
    </seg>
    <seg id="b.GEN.1.2" type="verse">
        En die aarde was woes en leeg, en duisternis was op die
    wêreldvloed, en die Gees van God het gesweef op die waters.
    </seg>
    ...
</div>
</div>
</body>
</text>
</cesDoc>
```

Slika 5.: primjer CES dokumenta s 1. razinom označavanja primarnih podataka[99]

[97] Ide, N. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. 1999.
[98] Corpus Encoding Standard. // Dostupno na: https://www.cs.vassar.edu/CES/CES1-0.html
[99] Christodouloupoulos, C.; Steedman, M. A massively parallel corpus: the Bible in 100 languages

U CES-u se razlikuju **primarni podaci**, odnosno „neoznačeni“ podaci u elektroničkom obliku (najčešće izvorno nisu namijenjeni za lingvističke svrhe, primjerice objavljivanje i emitiranje) te **lingvističke oznake**, koje čine informacije izrađene i dodane primarnim podacima kao rezultat neke lingvističke analize. CES obuhvaća kodiranje objekata u primarnim podacima koji se čine relevantni za zadatke temeljene na korpusima u istraživanju i primjeni jezičnog inženjeringa, uključujući:

1) Označavanje na razini dokumenta: bibliografski opis dokumenta, opis kodiranja itd.
2) Strukturno označavanje: strukturne jedinice teksta kao što su volumen, poglavlja itd. do razine paragrafa, također fusnote, naslovi, zaglavlja, tablice, slike itd.
3) Označavanje za strukture u okviru paragrafa: rečenice, citati, riječi, kratice, nazivi, datumi, pojmovi itd.[100]

Osim toga, za CES su definirane i tri razine kodiranja za primarne podatke, a primjer prve razine prikazan je na gornjoj slici. Više informacija o CES-u moguće je pronaći u potpunoj dokumentaciji na internetskoj stranici *https://www.cs.vassar.edu/CES/*.

XCES označava CES standard prebačen u XML okruženje. Temelji se na jednakoj arhitekturi podataka koja se sastoji od primarnog kodiranog teksta i udaljenih (eng. *standoff*) oznaka u zasebnim dokumentima.[101]

5.2.2. TEI

Inicijativa za kodiranje teksta (*Text Encoding Initiative,* TEI) je konzorcij koji kolektivno razvija i održava standard za prikaz tekstova u digitalnom obliku. TEI je razvio skup Smjernica koje određuju metode kodiranja za strojno čitljive tekstove, osobito u humanističkim i društvenim znanostima i lingvistici. Od 1994. godine TEI smjernice široko koriste knjižnice, muzeji, izdavači i pojedinačni znanstvenici za prikaz tekstova na internetu za istraživanje, učenje i očuvanje. TEI smjernice definiraju „shemu kodiranja“ prikazanu u formalnom jeziku za označavanje podataka. Izvorni TEI jezik (verzija P1 do P3) koristio je sintaksu SGML-a. S verzijom P4 uvedena je mogućnost izbora između SGML-a ili XML-a. Od verzije P5 dostupan

[100] Ide, N. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. 1999.
[101] Corpus Encoding Standard. // Dostupno na: https://www.cs.vassar.edu/CES/CES1-0.html

je isključivo kao XML. Osim toga, P5 uvelike se oslanja na XML standarde poput shematskih jezika i alata za programiranje kao što su XSLT i XQuery.[102]

Poput drugih jezik za označavanje, TEI jezik definira skup oznaka XML elemenata koji se koriste za kodiranje tekstova, zajedno s atributima koji služe za izmjenu elemenata. S obzirom da TEI smjernice nastoje pružiti okvir za kodiranje (u teoriji) bilo kojeg žanra teksta iz bilo kojeg razdoblja u bilo kojem jeziku, potpuni TEI skup oznaka vrlo je bogat i sastoji se od gotovo 500 elemenata. U praksi većina korisnika TEI-a rutinski upotrebljava mnogo manji podskup potpunog jezika. Elementi TEI skupa oznaka spadaju u dvije široke kategorije: prva služi za obilježavanje metapodataka o tekstu koji se kodira (autor, bibliografske informacije, povijest revizija itd.), dok druga služi za kodiranje strukturnih značajki samog dokumenta, kao što su odjeljci, zaglavlja, paragrafi, citati, istaknut tekst itd.[103]

S obzirom da se TEI kodiranje može primijeniti na mnogo različitih vrsta tekstova, osmišljeno je tako da bude vrlo modularno: korisnici mogu inkorporirati skupove značajki prilagođene za specifične žanrove, kao što su dramski tekstovi, stari rukopisi, prijepis govora, ispisani rječnici, kritike i mnogi drugi. U ranim verzijama Smjernica to se postizalo definiranjem potrebne „jezgre“ skupa oznaka, nekoliko „temeljnih“ skupova koji odgovaraju glavnim žanrovima (proza, lirika, drama itd.) i nekoliko „dodatnih“ skupova oznaka za specijalizirane značajke poput naziva, datuma ili lingvističkih analiza. Verzija P5 još je više modularna: zadržan je koncept jezgre s ključnim uobičajenim elementima, a svi ostali skupovi oznaka smatraju se dodatnim modulima koji se mogu kombinirati, izmjenjivati i smanjiti prema korisnikovim potrebama.[104]

TEI jezik također je osmišljen tako da bude proširiv. TEI smjernice opisuju postupke na koje korisnici mogu dodati, izmijeniti ili preimenovati elemente i atribute da bi odgovarali njihovim potrebama. Također je moguće inkorporirati druge XML jezike kao što su MathML ili RDF u dokument kodiran prema TEI shemi. Usklađenost s TEI standardom definirana je formalno na takav način da se većina može provjeriti s pomoću softvera. Iako se točna definicija mijenjala tijekom razvoja novih inačica TEI-a, kodirani dokument u suštini je usklađen s TEI standardom:

- Ako je dobro formatiran XML dokument

[102] Text Encoding Initiative. // Dostupno na: http://www.tei-c.org/index.xml
[103] Ibid.
[104] Ibid.

- Ako je potvrđen u skladu sa standardnom TEI shemom ili shemom koja je izrađena putem prilagodbi u skladu s dopuštenjima iz dokumentacije TEI Smjernica
- Ako su sve preinake TEI skupa oznaka pravilno dokumentirane, obično u datotekama korištenima za izradu prilagođene sheme.[105]

```
<body>
   <pb xml:id="leaf01r" type="recto"/>
   <lg type="poem">
      <head rend="underline" type="main-authorial">After <subst>
         <del type="overstrike" seq="1">an</del>
         <add place="supralinear" type="insertion" seq="2">the <del type="overstrike">unsolv'd</del> </add>
      </subst> argument</head>
      <l>
         <seg>
            <del type="overstrike">The</del>
            <add place="supralinear" type="insertion">
               <del type="overstrike">Coming in,</del>
               <subst>
                  <del type="overwrite" seq="1">a</del>
                  <add place="over" type="overwrite" seq="2">A </add>
               </subst>
            group of </add>
         little children, and their</seg>
         <seg>ways and chatter, flow
            <add place="inline" type="unmarked">in, </add>
            <del type="overstrike">
               <add place="supralinear" type="unmarked">upon me</add>
            </del>
         </seg>
      </l>
      <l>
         <seg>Like <add place="supralinear" type="insertion">welcome </add> rippling water o'er my </seg>
         <seg>heated <add place="supralinear" type="insertion">nerves and </add> flesh.</seg>
      </l>
      <closer>
         <signed>Walt Whitman</signed>
      </closer>
   </lg>
</body>
```

Slika 6.: primjer elementa *body* u XML datoteci prema TEI standardu[106]

Iako je korisno primijeniti općenite standarde anotacije poput prethodno navedenih za učeničke podatke u općenitom smislu, u području označavanja pogrešaka istraživači još uvijek nisu dogovorili standardnu općenitu shemu. Kategorizacija učeničkih pogrešaka je mukotrpan i često bezuspješan posao s obzirom da postoje razni načini klasifikacije pogrešaka, ovisno o interesima istraživača i uključenim teorijama te je klasifikacija često valjana samo za teoriju na kojoj je temeljena. Osim toga, većina ljudi ima različite poglede na vrste pogrešaka što može dovesti do vrlo niskog podudaranja među različitim osobama koje vrše klasifikaciju. S obzirom

[105] Text Encoding Initiative. // Dostupno na: http://www.tei-c.org/index.xml
[106] TEI By Example. // Dostupno na: http://teibyexample.org/TBE.htm

na navedeno, općeniti skup oznaka za pogreške čini se kao vrlo dobar cilj kojem treba težiti. Čak i u slučaju da općeniti ili standardni skup oznaka nije moguće u potpunosti iskoristiti za neke istraživačke svrhe, takav općeniti skup oznaka i dalje može biti dobra početna točka te se može prilagoditi i nadograditi za konkretne istraživačke probleme.[107]

5.3. Analiza prikupljenih podataka

Nakon prikupljanja i obrade podataka, sljedeći je korak sama analiza učeničkog korpusa. Prilikom uporabe učeničkih korpusa u lingvističke svrhe najčešće se koriste dva metodološka pristupa. To su kontrastivna analiza (međujezika) i (računalno potpomognuta) analiza pogrešaka.

5.3.1. Kontrastivna analiza međujezika

Kontrastivna analiza međujezika (*contrastive interlanguage analysis* – CIA) je, kao što sam naziv kaže, kontrastivna, što znači da se oslanja na usporedbu, a sastoji se od provođenja kvantitativnih i kvalitativnih usporedbi između jezika izvornih govornika (NS – *native speaker*) i neizvornih govornika (NNS – *non-native speaker*) ili između različitih varijeteta jezika neizvornih govornika.[108] Kontrastivna analiza uvelike se upotrebljavala u području usvajanja drugog jezika tijekom 60-ih i 70-ih godina prošlog stoljeća kao metoda kojom se pokušavalo objasniti zašto se određene značajke jezika usvajaju teže od drugih. S obzirom na tadašnje prevladavajuće biheviorističke teorije prema kojima je jezik pitanje habituacije, odnosno navika, težina ovladavanja drugim/stranim jezikom ovisila je o razlici između materinskog jezika i jezika koji se uči.[109] U samoj kontrastivnoj analizi međujezika možemo govoriti o dvije vrste usporedbe. U prvom slučaju radi se o usporedbama NS/NNS čiji je cilj dati bolji uvid u značajke pisanog i govorenog jezika učenika kroz detaljne usporedbe lingvističkih značajki jezika u korpusima izvornih i neizvornih govornika. Na taj način moguće je uz pogreške otkriti i širok raspon značajki u pisanom i govornom jeziku učenika koje otkrivaju da se radi o neizvornom govorniku, primjerice pretjeranu ili premalenu uporabu određenih riječi, fraza i sintaktičkih konstrukcija. Iako neki lingvisti odbacuju takve usporedbe jer smatraju da se međujezik treba proučavati kao zasebni sustav, praktična primjena učeničkih korpusa

[107] Tono, Y. Learner corpora: design, development and applications. 2003.
[108] Granger, S. A Bird's-eye view of learner corpus research, 2002.
[109] Contrastive analysis. // Wikipedia: the free encyclopedia. (19.7.2017.)

uvjetovana je usporedbom s jezikom izvornih govornika s obzirom da je cilj poučavanja stranog jezika uvijek poboljšavanje učenikove vještine u uporabi stranog jezika, što u suštini znači približavanje određenim normama „izvornoga“ jezika.[110]

U sklopu kontrastivne analize također se nalaze usporedbe NNS/NNS, odnosno usporedbe jezika dvije skupine neizvornih govornika koje proširuju saznanja o međujeziku. Osobito su korisne usporedbe učenika s različitim materinskim jezicima jer pomažu istraživačima razlikovati značajke koje su zajedničke za više skupina učenika te su stoga vjerojatnije povezane s razinom poznavanja stranog jezika od značajki koje su posebne za učenike s određenim materinskim jezikom te su stoga potencijalno povezane s tim konkretnim jezikom.[111]

5.3.2. Računalno potpomognuta analiza pogrešaka

Računalno potpomognuta analiza pogrešaka usmjerena je na pogreške u međujeziku te se služi računalima za označavanje, dohvaćanje i analiziranje istih.[112] Analizu pogrešaka u području usvajanja drugog jezika ustanovio je 1960-ih Stephen Pit Corder. Bila je to alternativa kontrastivnoj analizi koja je dokazala da kontrastivna analiza ne može predvidjeti veliku većinu pogrešaka, a ključno otkriće analize pogrešaka jest da mnogi učenici rade pogreške jer donose krive zaključke o pravilima novog jezika. Istraživači u području analize pogrešaka razlikuju pogreške, odnosno sustavnu krivu uporabu jezika, i propuste, nenamjernu krivu uporabu jezika unatoč poznavanju pravila. Još jedna bitna značajka analize pogrešaka jest pokušaj klasifikacije pogrešaka prema raznim kriterijima. Metodološki problemi bili su prisutni od samog početka s obzirom da je prema isključivo lingvističkim podacima često bilo nemoguće pouzdano odrediti vrstu pogreške. Isto tako, analiza pogrešaka ograničena je isključivo na učenikovu proizvodnju (govor, pisanje) te ne može obuhvatiti njegovo razumijevanje jezika (slušanje, čitanje). Dodatni problem predstavljaju komunikacijske strategije poput „izbjegavanja“, pri čemu učenici jednostavno ne upotrebljavaju oblike i konstrukcije za čiji način uporabe nisu sigurni. Zbog navedenih razloga naposljetku je odbačena sveobuhvatna teorija o učeničkim pogreškama, a

[110] Granger, S. A Bird's-eye view of learner corpus research, 2002.
[111] Granger, S. A Bird's-eye view of learner corpus research, 2002.
[112] Ibid.

analiza pogrešaka nastavila se upotrebljavati za neke konkretne probleme u usvajanju drugog jezika.[113]

Pristupi učeničkim korpusima orijentirani na pogreške znatno se razlikuju od prijašnjih istraživanja analize pogrešaka s obzirom da su potpomognuti računalima i uključuju višu razinu standardizacije, a možda još važnije od toga – pogreške su prikazane u potpunom kontekstu teksta uz ispravne oblike. U računalno potpomognutoj analizi pogrešaka obično se koriste dvije metode. Prva metoda sastoji se od jednostavnog izdvajanja problematične lingvističke stavke (riječ, fraza, sintaktička konstrukcija) i pretraživanja korpusa s ciljem dohvaćanja svih pojava neispravne uporabe te stavke s pomoću standardnog softvera za dohvaćanje i pretraživanje teksta. Prednost je svakako brzina, ali nedostatak je ograničavanje na isključivo prethodno određene problematične stavke. Druga je metoda znatno sporija, ali je sveobuhvatnija i omogućava pronalaženje problema u učenju kojih istraživač možda nije bio svjestan. Radi se o razvijanju standardiziranog sustava oznaka za pogreške i označavanju svih pogrešaka u učeničkom korpusu. Iako je utrošak vremena za takvo označavanje poprilično velik, moguće ga je ubrzati određenim alatima za pogreške, a nakon dovršetka označavanja pogrešaka takav korpus pruža ogromne istraživačke mogućnosti. Bitno je naglasiti da, iako neki smatraju da koncentriranje na pogreške nije pozitivan način učenja/poučavanja jezika, analiza pogrešaka ima svoju ulogu u postupku razumijevanja razvoja međujezika i u pedagoškim okvirima može učiteljima omogućiti bolju predodžbu razine znanja učenika i djelomično usmjeriti poučavanje na problematična područja.[114]

[113] Error analysis. // Wikipedia: the free encyclopedia. (19.7.2017.)
[114] Granger, S. A Bird's-eye view of learner corpus research, 2002.

5.4. Sketch Engine

Sketch Engine je alat za upravljanje i analizu korpusa koji je 2003. godine izradio Adam Kilgarriff, odnosno njegova tvrtka Lexical Computing Limited.[115] Danas je to vodeći alat takve vrste koji sadržava oko 400 tekstualnih korpusa sa ukupno 20 milijardi riječi na 90 različitih jezika.[116] Iako je Sketch Engine komercijalni softver, glavne značajke temeljnih alata Manatee i Bonito koje su razvijene do 2003. (i kasnije nadograđivane) dostupne su u besplatno u sklopu paketa NoSketchEngine pod licencijom GPL. U smislu arhitekture, Sketch Engine se sastoji od tri glavne komponente: temeljnog sustava za upravljanje bazom podataka Manatee, internetskog korisničkog sučelja za pretraživanje Bonito te internetskog sučelja za izradu i upravljanje korpusima pod nazivom Corpus Architect.[117]

Glavne funkcije alata Sketch Engine su sljedeće:

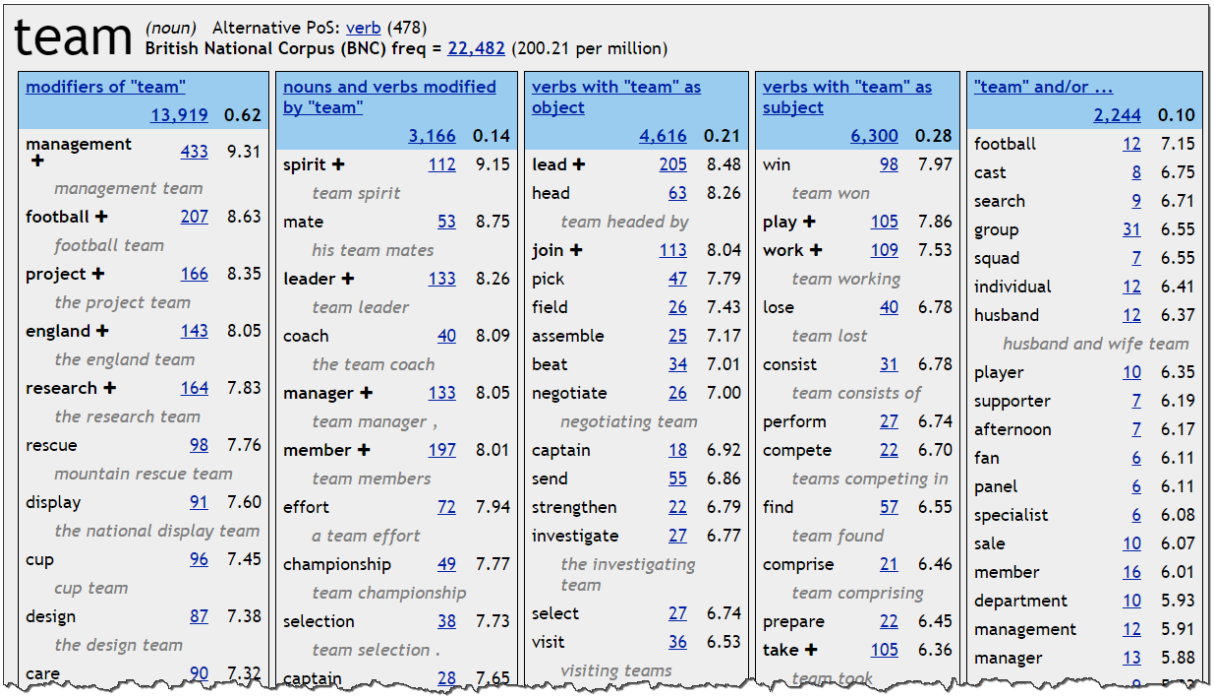
- Značajka „skica riječi“ (eng. *word sketch*) po kojoj je alat i dobio naziv, a koja daje automatske popise veličine jedne stranice o gramatičkom i kolokacijskom okruženju riječi dobivene iz korpusa
- Program za konkordance koji omogućava upite s proširenom sintaksom CQL-a (jezik za korpusne upite, eng. *corpus query language*), ali nudi i pojednostavljeno i intuitivno sučelje za korisnike koji nisu upoznati s CQL-om
- Značajka popisa riječi koja omogućava izrade popisa riječi iz korpusa prema različitim kriterijima, uključujući izvlačenje ključnih riječi iz podkorpusa i također omogućava izrade frekvencijskih popisa metapodataka (poput domene dokumenta)
- Statistički tezaurus temeljen na kolokacijama dobivenim iz „skica riječi“[118]

Sve navedene funkcije dostupne su na stranici *https://www.sketchengine.co.uk* zajedno s prethodno učitanim korpusima i sustavom za izradu vlastitog korpusa (uz plaćanje, kako je prethodno navedeno).

S obzirom na zahtjeve korisnika, u Sketch Engine s vremenom je dodana i nova funkcija koja olakšava rad s učeničkim korpusima u kojima su označene pogreške i ispravci. U nastavku će

[115] Sketch Engine. // Wikipedia: the free encyclopedia (28.9.2017.)
[116] Sketch Engine. // Dostupno na: https://www.sketchengine.co.uk/
[117] Sketch Engine. // Wikipedia: the free encyclopedia (28.9.2017.)
[118] Kovář, V.; McCarthy, D. New Learner Corpus Functionality in the Sketch Engine, 2012.

biti opisan postupak izrade učeničkog korpusa i označavanja pogrešaka/ispravaka u alatu Sketch Engine.



team *(noun)* Alternative PoS: verb (478)
British National Corpus (BNC) freq = 22,482 (200.21 per million)

modifiers of "team"			nouns and verbs modified by "team"			verbs with "team" as object			verbs with "team" as subject			"team" and/or ...		
	13,919	0.62		3,166	0.14		4,616	0.21		6,300	0.28		2,244	0.10
management +	433	9.31	spirit +	112	9.15	lead +	205	8.48	win	98	7.97	football	12	7.15
management team			*team spirit*			head	63	8.26	*team won*			cast	8	6.75
football +	207	8.63	mate	53	8.75	*team headed by*			play +	105	7.86	search	9	6.71
football team			*his team mates*			join +	113	8.04	work +	109	7.53	group	31	6.55
project +	166	8.35	leader +	133	8.26	pick	47	7.79	*team working*			squad	7	6.55
the project team			*team leader*			field	26	7.43	lose	40	6.78	individual	12	6.41
england +	143	8.05	coach	40	8.09	assemble	25	7.17	*team lost*			husband	12	6.37
the england team			*the team coach*			beat	34	7.01	consist	31	6.78	*husband and wife team*		
research +	164	7.83	manager +	133	8.05	negotiate	26	7.00	*team consists of*			player	10	6.35
the research team			*team manager ,*			*negotiating team*			perform	27	6.74	supporter	7	6.19
rescue	98	7.76	member +	197	8.01	captain	18	6.92	compete	22	6.70	afternoon	7	6.17
mountain rescue team			*team members*			send	55	6.86	*teams competing in*			fan	6	6.11
display	91	7.60	effort	72	7.94	strengthen	22	6.79	find	57	6.55	panel	6	6.11
the national display team			*a team effort*			investigate	27	6.77	*team found*			specialist	6	6.08
cup	96	7.45	championship	49	7.77	*the investigating team*			comprise	21	6.46	sale	10	6.07
cup team			*team championship*			select	27	6.74	*team comprising*			member	16	6.01
design	87	7.38	selection	38	7.73	visit	36	6.53	prepare	22	6.45	department	10	5.93
the design team			*team selection .*			*visiting teams*			take +	105	6.36	management	12	5.91
care	90	7.32	captain	28	7.65				*team took*			manager	13	5.88

Slika 7.: prikaz funkcije skica riječi za englesku riječ *team*[119]

5.4.1. Izrada učeničkog korpusa

Sketch Engine ima uz mogućnost učitavanja vlastitih datoteka i jedinstvenu mogućnost izrade korpusa od relevantnih tekstova s interneta. S obzirom da se radi o učeničkom korpusu, jasno je da je potrebna prva opcija, odnosno učitavanje vlastitih datoteka. Prije učitavanja samih datoteka, potrebno je upisati naziv korpusa i odabrati jezik na kojem će biti tekst iz datoteka. Ukoliko želite upotrebljavati funkciju „skica riječi“, potrebno je učitati i gramatiku za taj jezik u odgovarajućem obliku, a moguće je upotrijebiti već postavljenu gramatiku za taj jezik ukoliko postoji. Preporučeno je sve metapodatke (uključujući naziv dokumenta) upisati u obliku XML struktura. Sketch Engine prima dokumente u nizu različitih formata, uključujući, među ostalima, .doc, .docx, .htm, .html, .pdf, .txt i .xml. Prije uporabe korpusa potrebno ga je i kompilirati. Kompiliranje uključuje primjenu nekoliko alata koji obrađuju podatke iz korpusa u svrhu omogućavanja svih funkcija alata, uključujući izradu skica riječi, tezaurusa, n-grama i trendova.[120]

[119] Sketch Engine. // Dostupno na: https://www.sketchengine.co.uk/
[120] Ibid.

Sketch Engine obično zahtijeva vertikalni format teksta (jedna riječ po redu, s mogućnosti više stupaca za leme, oznake itd. i strukturne oznake u obliku XML-a). Pogreške se označavaju oznakom <err>, a ispravci oznakom <corr> te su te oznake i njihovo pravilno zatvaranje (</err> i </corr>) obavezne u korpusu s označenim pogreškama. Ispravak mora uvijek slijediti odmah nakon pogreške.[121]

```
We          we          PR
learn       learn       VV
maths       math        NN
to          to          PP
<err type="BadWording">
<err type="Typo">
caan        caan        ??
</err>
<corr type="Typo">
can         can         VA
</corr>
</err>
<corr type="BadWording">
be          be          VB
able        able        VP
to          to          PP
</corr>
compute     compute     VV
our         our         PR
taxes       tax         NN
```

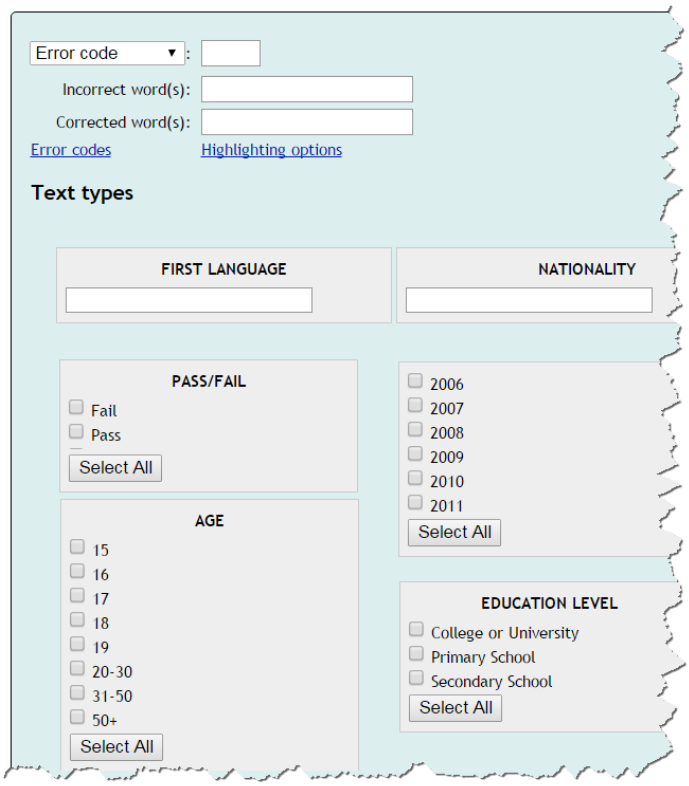
Slika 8.: Primjer teksta iz učeničkog korpusa[122]

Osim toga, vrsta pogreške može se navesti s pomoću atributa *type*, a vrijednost atributa mora biti jednaka u oznaci za pogrešku i ispravak. Oznake za pogrešku i ispravak mogu biti prazne, ali u tom slučaju potrebno je upisati posebnu pojavnicu „===NONE===“.[123] Na slici 8. vidljivo je kako označavanje teksta u vertikalnom obliku izgleda u praksi. U lijevom stupcu nalazi se izvorni tekst, srednji stupac sadrži leme riječi, dok se u posljednjem stupcu nalazi podatak o vrsti riječi. Oznake za pogreške i ispravke upisane su kako je prethodno navedeno u XML strukturama.

Sketch Engine ima namjensko sučelje za rad s učeničkim korpusima (slika 9.). To sučelje omogućava korisniku pretraživanje prema samim pogreškama, vrsti pogreške, ispravku pogreške ili prema kombinaciji bilo kojih navedenih kriterija. Osim toga, bilo koji metapodaci navedeni u korpusu mogu se upotrijebiti u pretrazi i analizirati radi dobivanja informacija o

[121] Sketch Engine. // Dostupno na: https://www.sketchengine.co.uk
[122] Kovář, V.; McCarthy, D. New Learner Corpus Functionality in the Sketch Engine, 2012.
[123] Sketch Engine. // Dostupno na: https://www.sketchengine.co.uk

primjerice distribuciji učeničkih pogrešaka po dobnim skupinama, razini poznavanja jezika, materinskom jeziku, vrsti zadatka i sl. Mogućnosti pretraživanja u principu su ograničene isključivo količinom, odnosno opsegom označavanja metapodataka u tekstovima. Svi metapodaci uneseni za neki tekst mogu predstavljati kriterij za pretraživanje.[124]



Slika 9.: prikaz sučelja za učeničke korpuse[125]

Nakon pretraživanja učeničkog korpusa pogreške u rezultatima obično se prikazuju crvenom bojom, dok su ispravci prikazani zelenom bojom radi bolje uočljivosti. Način prikaza moguće je promijeniti definiranjem stavke DISPLAYCLASS u konfiguracijskoj datoteci za dvije odgovarajuće definicije struktura (lijeva tablica). Zatim se u CSS datoteci (view.css) mogu definirati klase i dodati stilovi po želji (desna tablica).[126]

```
STRUCTURE err {
    DISPLAYCLASS "errclass"
}
STRUCTURE corr {
    DISPLAYCLASS "corrclass"
}
```

```
.errclass {
    background-color: red;
    color: white;
    font-weight: bold;
}
.corrclass {
    background-color: green;
    color: white;
    font-weight: bold;
}
```

[124] Sketch Engine. // Dostupno na: https://www.sketchengine.co.uk
[125] Ibid.
[126] Ibid.

Sličan prijedlog prikaza pogrešaka i ispravaka u učeničkim korpusima dali su Kosem i ostali u svojem radu „The Sketch Engine interface for a learner corpus annotated with errors and corrections“. U njihovoj preinaci sučelja za korpus Šolar prikaz pogrešaka i ispravaka (bez oznaka) predstavlja standardnu opciju, a obje vrste informacije jasno se razlikuju jedna od druge te okolnog teksta koji ne čini pogreške/ispravke. Pogreške su prikazane crvenom, a ispravci zelenom bojom, dok je sav ostali tekst crne boje. Osim toga, crvena boja (standardna postavka) za osnovnu riječ zamijenjena je crnom (podebljano) što znači da korisnici mogu jednostavno razlikovati pogreške i ostali tekst prilikom provođenja jednostavnog pretraživanja (ali ne i kod pretraživanja pogreški). Takvo sučelje za korpus uvelike pomaže korisnicima koji nisu stručnjaci u uporabi korpusa.[127]

oma brez učinka, saj sta za ljubezen potrebna dva. Brez	nobenega	| vsakega truda se ne da priti nikamor, vendar Ofelija tega žal
| , da odidejo , | oziramo da se vdajo. Na koncu ga	noben	| nihče ne zapusti in se skupaj bojujejo proti sovražniku. Pogu
kar se dogaja po svetu v določenih državah ni človeško,	noben	| nihče si ne zasluži takšnega obnašanja, kakršnega so deležni
Juliji | Julijo | , ki je mrtva ležala | ležala mrtva in	noben	| nihče mu ni povedal | , da v resnici ni mrtva. Ko se
človeka se lahko spremenimo v junaka, ki je glavni in mu	noben	| nihče nič ne more. V našem stvarnem življenju imamo ljudje
jegovo sobo, ki jo je imel v kleti, kajti tja ni smel hoditi	noben	| nihče samo | razen on | njega . Izidor kazen sprejme | je kaz
| , kajti v današnjih odnosih je čisto drugače | .	nobeden | Noben	| Noben oče nebi | ne bi svojemu sinu odrezal prst, ker naj bi
ekaterim "urejenim" družinam dogajajo pretepi in noče	noben	| nihče priznati tega. Nasilja ne podpira nihče.
biti njegovo razmišljanje | , medtem , | ko mu	noben	| nihče ne verjame, temveč ga vsi le obsojajo. Tega si noben |
n | nihče ne verjame, temveč ga vsi le obsojajo. Tega si	noben	| nihče izmed nas ne mora predstavljati, saj še nismo bili v

Slika 10.: standardni prikaz sučelja za korpus Šolar (prikazane su pogreške i ispravci)[128]

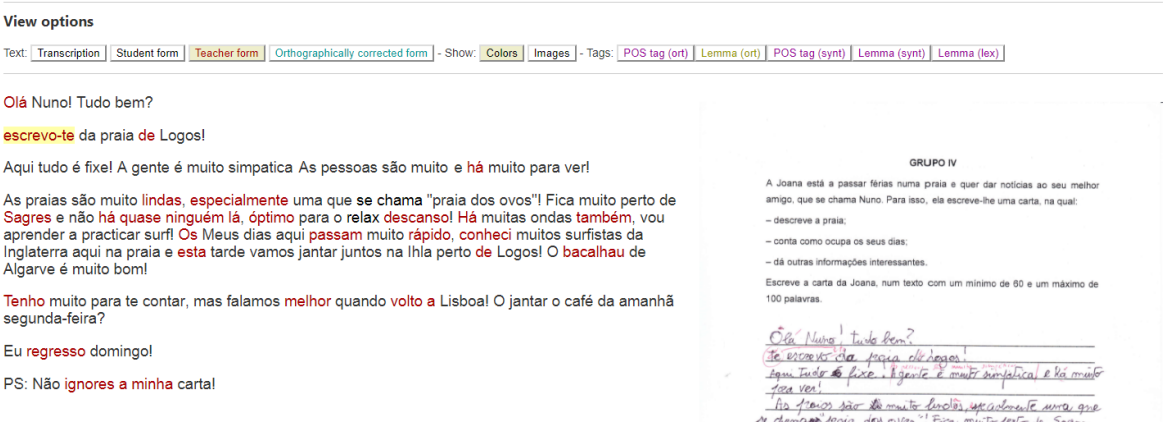
5.5. TEITOK

TEITOK je mrežni sustav za izradu, anotaciju i distribuciju korpusa. Nastao je kao rješenje za problem nemogućnosti zadržavanja tekstualne anotacije u većini lingvistički označenih korpusa zbog nedostatka odgovarajućih alata. Glavni cilj sustava TEITOK upravo je ponuditi rješenje koje omogućava dodavanje lingvističke anotacije na već tekstualno označene tekstove. Sustav omogućuje jednostavan prikaz XML datoteka, uređivanje metapodataka i pojedinačnih pojavnica te pretraživanje korpusa.[129]

Korpus u TEITOKU sastoji se od XML datoteka izrađenih prema TEI standardu uz malu modifikaciju u vidu tokenizacije. Tokenizacija se u sustavu TEITOK dodaje unutar teksta u TEI dokument. Od TEI standarda razlikuje se po tome što se pojavnice označavaju elementom

[127] Kosem, I.; Kovar, V.; Baisa, V.; Kilgarriff, A. The Sketch Engine interface for a learner corpus annotated with errors and corrections
[128] Ibid.
[129] Janssen, M. TEITOK: Text-Faithful Annotated Corpora, 2016.

<tok>, a sve lingvističke oznake upisuju se kao atributi u te elemente postavljene oko riječi. Osim toga, ne upotrebljava se element <choice> nego je u elementu <tok> moguće kao atribute upisati različite ortografske oblike iste riječi (ako postoje). Obilježeni tekst prikazuje se izravno u pregledniku. Drugim riječima, XML *body* element teksta umetnut je u HTML stranicu, a u CSS-u specifičnom za projekt definirano je kako se prikazuju različiti XML elementi u tekstu – primjerice element <del> (izbrisana riječ) prikazan je u sivoj boji i precrtan. Također je moguće pokraj teksta prikazati i sliku (npr. skenirani tekst učenika). Kada postoje različite ortografske verzije teksta (izvorna verzija, standardiziran tekst, ispravci nastavnika i sl.), moguće je promijeniti verziju koja se prikazuje s pomoću tipki iznad teksta, kako je prikazano na slici 11.[130]



Slika 11.: prikaz datoteke s ispravcima nastavnika u crvenoj boji u sustavu TEITOK[131]

Radi lakšeg rada s metapodacima, TEITOK pruža mogućnost definiranja skupa metapodataka koji su relevantni za projekt i izradu tablice za uređivanje prema tim podacima. Tablica za uređivanje je HTML tablica koja opisuje sva relevantna polja uz koja se nalaze XPath definicije koje označavaju određeno polje u TEI zaglavlju (teiHeader). TEITOK s pomoću te tablice izrađuje jednostavan HTML obrazac koji zamjenjuje sve XPath poveznice s HTML poljima za upis te pretražuje u XML-u odgovarajuće vrijednosti i omogućava korisniku dodavanje ili izmjenu podataka. Nakon spremanja informacije se upisuju natrag u XML datoteku na odgovarajuće mjesto, pri čemu se izrađuju do tada nepostojeći čvorovi.[132]

S obzirom da je TEITOK zamišljen tako da bude što jednostavniji za rad, sustav pruža mogućnost primjene raznih vrsta obrade TEI dokumenata uz jedan klik mišem. To obuhvaća

[130] Janssen, M. TEITOK: Text-Faithful Annotated Corpora, 2016.
[131] Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2) // Dostupno na: http://teitok.iltec.pt/peapl2/index.php?action=home (28.9.2017.)
[132] Janssen, M. TEITOK: Text-Faithful Annotated Corpora, 2016.

tokenizaciju, označavanje vrsta riječi, lematizaciju i izradu CQP korpusa. Za označavanje vrsta riječi TEITOK upotrebljava NeoTag označivač neovisan o jeziku. Osim označavanja teksta u korpusu, NeoTag može upotrijebiti i sam korpus kao korpus za učenje da bi se dobio označivač koji je specijaliziran za određenu vrstu tekstova. Označavanje se vrši izravno u XML datoteku. Iako je moguće izravno pretraživati XML datoteke, radi jednostavnosti i brzine CQP verzija korpusa izrađuje se automatski, a upiti se mogu upisivati s pomoću CQL-a izravno na internetskoj stranici. Upit se obrađuje u CQP pokretaču, a rezultati se prikazuju u pregledniku u obliku KWIC reda za svaki rezultat. Pri izvozu u CQP potrebno je odrediti koji ortografski oblici će se preuzeti. Iako je CQP namijenjen za lingvističke upite u korpusu, sučelje također omogućava pretraživanje dokumenata. Ukoliko se ne odabere niti jedan upit povezan s pojavnicom i zatraže se isključivo dokumenti prema određenom datumu, mjestu, materinskom jeziku ili nekom drugom metapodatku, rezultati neće biti KWIC popis ili prikaz konteksta nego popis dokumenata s traženim karakteristikama metapodataka.[133]

Slika 12.: primjer sučelja u alatu TEITOK (korpus PEAPL2)[134]

Uređivanje lingvističkih oznaka vrlo je jednostavno u TEITOKU: u tekstualnom prikazu teksta sve oznake za riječ prikazuju se u skočnom prozoru kada se pokazivač postavi na tu riječ. Bilo koja pojavnica koju treba ispraviti može se jednostavno urediti pritiskom mišem na tu riječ, pri čemu će se otvoriti HTML obrazac s prikazom sadržaja pojavnice. Pritiskom na spremi izravno će se ažurirati sadržaj pojavnice u XML datoteci. Stoga je primjerice prilikom normalizacije

[133] Janssen, M. TEITOK: Text-Faithful Annotated Corpora, 2016.
[134] Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2) // Dostupno na: http://teitok.iltec.pt/peapl2/index.php?action=home (28.9.2017.)

teksta potrebno jednostavno pročitati tekst i ispraviti sve pojavnice koje još nisu normalizirane.[135]

5.5.1. Izrada učeničkog korpusa (COPLE2)

Korpus COPLE2 pisani je i govorni učenički korpus portugalskog jezika. Pisani se dio sastoji od 966 slobodnih eseja koje su napisala 424 studenta u razdoblju od 2010. i 2012. godine u dvije različite situacije: kao redovne ispite na predavanjima portugalskog kao stranog jezika na Institutu za portugalski jezik i kulturu (Instituto da Cultura e Língua Portuguesa, ICLP) i kao akreditacijski ispit u Centru za procjenu znanja portugalskog kao stranog jezika (Centro de Avaliação de Português Língua Estrangeira, CAPLE) pri Fakultetu društvenih i humanističkih znanosti Sveučilišta u Lisabonu (FLUL). Za korpus su odabrani materinski jezici koji su imali minimalno šest tekstova u početnom skupu podataka. Tekstovi su podijeljeni u pet kategorija poznavanja jezika prema ZEROJ-u: A1, A2, B1, B2 i C1. Studenti su radili različite zadatke koji spadaju u sljedeće kategorije: dijalog, formalno i osobno pismo, informativni esej, poruka/e-mail, raspravljački esej, prepričavanje, recenzija knjige i slično (raspravljački eseji čine 36 % tekstova).[136]

Metapodaci

Detaljni metapodaci kodirani su u svakoj datoteci u zaglavlju usklađenom s TEI standardom u XML formatu. Profil kandidata opisan je u 20 polja, dok je zadatak i tekst opisan u 14. Metapodaci o sudionicama dobiveni su iz registracijskog obrasca za kolegij Portugalski kao strani jezik. S obzirom da se obrazac ispunjavao na papiru, mnogi su podaci često nedostajali pa je prema tome uspostavljeno 7 polja obaveznih za uključivanje sudionika u korpus COPLE2: ime, dob, nacionalnost, spol, materinski jezik, poznavanje dodatnih drugih ili stranih jezika, trajanje učenja portugalskog jezika. Poznavanje drugih jezika je izbačeno zbog čestog nedostatka podataka. Taj je problem naknadno riješen uvođenjem obaveznog *online* registracijskog sustava koji zahtijeva potpuno ispunjavanje potrebnih polja o svojem učeničkom profilu (primjerice informacije o razini poznavanja jezika, boravku u zemljama portugalskog govornog područja (gdje, kada, koliko dugo), obrazovanju, materinskom jeziku, jeziku kojim se služi kod kuće). Podaci o tekstu uključuju žanr, temu, opis zadataka (inicijalni

[135] Janssen, M. TEITOK: Text-Faithful Annotated Corpora, 2016.
[136] Mendes, A.; Antunes, S.; Janssen, M.; Gonçalves, A. The COPLE2 Corpus: a Learner Corpus for Portuguese. 2016.

ispit, polugodišnji ili završni ispit, domaća zadaća, akreditacijski ispit u CAPLE-u), postojanje vremenskog ograničenja, dostupnost referentnih materijala, broj pojavnica i datum.[137]

Imenovanje datoteka omogućuje identifikaciju ispitanika, razinu poznavanja jezika te vrstu ispita s pomoću dosljednog uzorka:

- 5 znakova označava kod ispitanika: 2 slova odnose se na prvi jezik (ISO 639-16) a 3 broja označavaju pojedinog ispitanika,
- vrsta kolegija: godišnji kolegij (CA) ili ljetni kolegij (CV),
- razina poznavanja jezika: pripremna (I), temeljna (E), prijelazna (M), samostalna (A) i napredna (S),
- vrsta ispita: inicijalni (TD), polugodišnji (TT) ili završni (TF).

Primjerice fr010CVITD označava ispitanika broj 10 kojemu je materinski jezik francuski, pohađao je ljetni kolegij (CV), razina poznavanja jezika je pripremna te je tekst napisao na inicijalnom ispitu (TD).[138]

Priprema podataka

Rukom pisani eseji prvo su skenirani i pohranjeni u .pdf formatu, a zatim su ručno prepisani. Rukopis studenata predstavljao je dva problema: onemogućio je automatsku digitalizaciju dokumenata s pomoću OCR-a te je prilikom prepisivanja teksta problem bio ispravno prepoznati riječ/slovo zbog sličnosti nekih slova. Prijepis je vrlo sličan izvornom dokumentu: zadržane se sve izmjene koje je student napravio tijekom postupka pisanja, kao što su brisanja i umetanja riječi ili promjena položaja segmenata te su iste kodirane u TEI sukladnoj XML datoteci. Isto tako zabilježene su i sve izmjene koje je napravio nastavnik. Prijelomi redova i riječi nisu zadržani, ali paragrafi jesu. Uklonjene su sve osobne informacije, kao što su imena, adrese, telefonski brojevi i zamijenjene s primjerice „XX“. Za svaki tekst dostupna je i pročišćena verzija u .txt formatu koja odgovara konačnoj verziji koju je zamislio student.[139]

Izbrisane riječi/segmenti označeni su elementom <del>, a umetanja elementom <add>. U oba elementa može se nalaziti atribut *hand* koji označava autora izmjene (student ili nastavnik). Bilo kakve oznake napravljene iznad riječi ili segmenta teksta kodirane su elementom <hi>, a dodatno se mogu opisati prema autoru (atribut *hand*) i vrsti oznake s pomoću atributa *rend*

[137] Mendes, A.; Antunes, S.; Janssen, M.; Gonçalves, A. The COPLE2 Corpus: a Learner Corpus for Portuguese. 2016.
[138] Ibid.
[139] Ibid.

(podcrtano, zaokruženo, prekriženo). Primjer navedenih oznaka i pročišćenog konačnog teksta vidljiv je na slici 13.[140]

(1) Normalmento, Eu acordo às oito horias de manhã, <del hand="zh010">t</del> e tomo o duche e o pequeno-almoço. Eu saio de casa e apanho o metro para universidade, eu chego o escritório de XX <del hand="corrector">á</del> <add hand="corrector">às</add> nove de manhã. <hi hand="corrector" rend="underlined">Eu escrevo <add hand="zh010">os</add></hi> livros de engenheiro, ou tenho curso.

(2) Normalmento, Eu acordo às oito horias de manhã, e tomo o duche e o pequeno-almoço. Eu saio de casa e apanho o metro para universidade, eu chego o escritório de XX á nove de manhã. Eu escrevo os livros de engenheiro, ou tenho curso.

Translation: Usually, I wake up at eight o'clock in the morning, and I take a shower and breakfast. I leave home and catch the metro to the university, I arrive to the office of XX at nine in the morning. I write the engineer books, or I have class.

Slika 13.: XML verzija (1) i pročišćena verzija (2) teksta[141]

Označavanje, vizualizacija i upiti

Nakon dovršetka prepisivanja, XML dokumenti uvezeni su u TEITOK radi vizualizacije, označavanja i pretraživanja. U TEITOKU je omogućeno interpretiranje XML kodiranja tako da je dostupan prikaz više različitih verzija pisanog teksta: prijepis (vizualizacija što bliža izgledu izvornog dokumenta), studentova verzija (završna verzija kako ju je zamislio student), ispravljena verzija (s označenim ispravcima koje je unio nastavnik) te slika ručno pisanog eseja (na zahtjev). Korpus je automatski tokeniziran te se za svaku pojavnicu može dodati normalizirana verzija u obliku atributa. Automatsko označavanje vrsta riječi i lematizacija napravljeni su u TEITOK okruženju s pomoću NeoTag označivača, koji je obučen na Referentnom korpusu suvremenog portugalskog jezika (CRPC).[142]

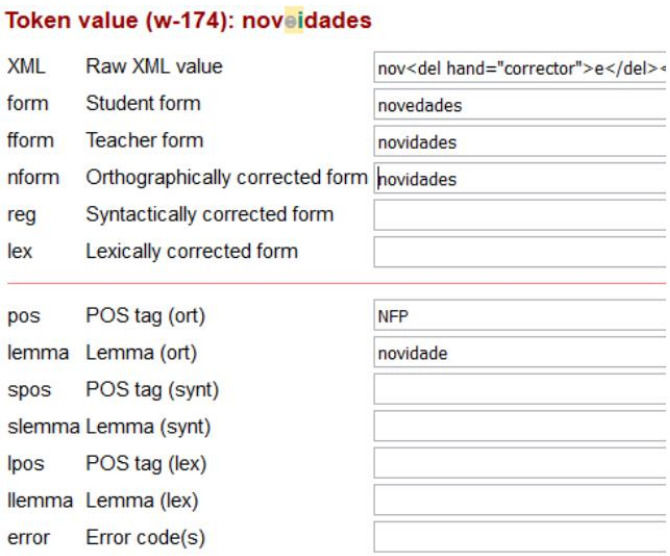
Pretraživanje je omogućeno s pomoću CQP-a. Kao što je prethodno navedeno, u izradi CQP korpusa izvoze se razne informacije: podaci na razini teksta poput informacija o studentu, informacije na razini pojavnice uključujući vrstu riječi, lemu, izvorni ortografski i normalizirani oblik te na razini segmenta informacije poput anotacije pogrešaka. U pretraživanju je moguće kombinirati različite vrste informacija, što omogućuje izvođenje složenih upita. Primjerice za portugalski jezik, usporedbom pisanog oblika problematičnog nastavka *–ão* s normaliziranim

[140] Mendes, A.; Antunes, S.; Janssen, M.; Gonçalves, A. The COPLE2 Corpus: a Learner Corpus for Portuguese. 2016.
[141] Ibid.
[142] Ibid.

oblikom moguće je pretražiti sve riječi koje su trebale biti napisane s nastavkom *–ão*, ali su napisane s primjerice *–am* (identičan izgovor) ili *–ao* (vizualno slično), a zatim dobiti distribuciju u tekstovima učenika istog materinskog jezika i procijeniti jesu li te pogreške specifične za određenu skupinu izvornih govornika.[143]

Anotacija pogrešaka

Za anotaciju pogrešaka odabrane su tri lingvističke razine anotacije: ortografska, gramatička i leksička. U svim slučajevima anotacija se sastoji od dodavanja ispravnog oblika riječi i njezine leme te vrste riječi. Sve tri razine za određenu se riječ mogu dodati u isto vrijeme. Na svakoj razini dolazi do različitih preinaka da bi se dobio ekvivalent jezika izvornog govornika. Prva se razina upotrebljava ako se u učenikovom tekstu nalazi pravopisna pogreška. Unosi se ortografski ispravljen oblik (*nform*), kao i odgovarajuća vrsta riječi (*pos*) te lema (*lemma*). Intervencije na ovoj razini uključuju oblike riječi i interpunkcijske znakove. Rješavaju se pogreške u interpunkciji, pravopisu i granicama riječi s ciljem dobivanja oblika iz jezika izvornog govornika koji je najsličniji učenikovom.[144]

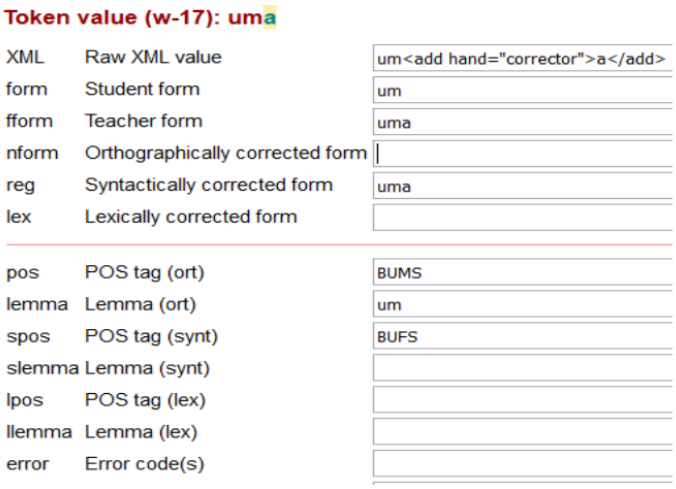


Slika 14: primjer ispravka ortografske pogreške[145]

Druga razina postaje potrebna ako postoji gramatička pogreška, odnosno ako je zbog riječi koju je upotrijebio student rečenica gramatički neispravna. Unosi se sintaktički ispravljen oblik (*reg*), kao i pripadajuća vrsta riječi (*spos*). Intervencije na gramatičkoj razini odnose se na

[143] Mendes, A.; Antunes, S.; Janssen, M.; Gonçalves, A. The COPLE2 Corpus: a Learner Corpus for Portuguese. 2016.
[144] del Rio, I.; Antunes, S.; Mendes, A.; Janssen, M. Towards error annotation in a learner corpus of Portuguese, 2016.
[145] Ibid.

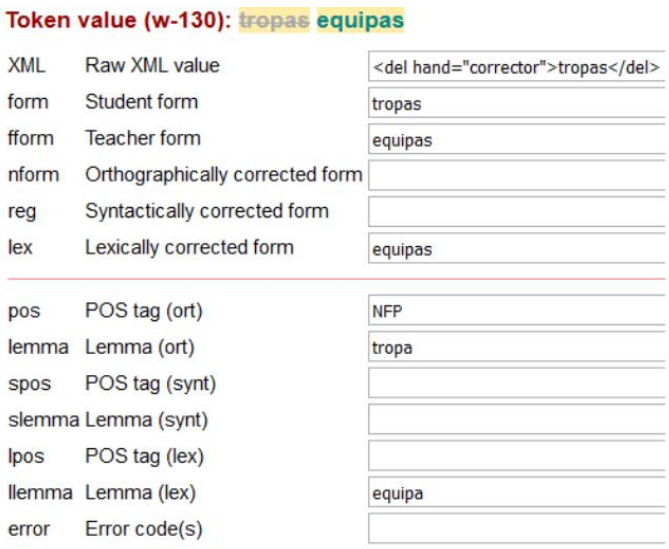
gramatičke probleme, odnosno pogreške koje nadilaze samu riječ i utječu na sintaktičke strukture. Iz tog razloga označivač mora uzeti u obzir kontekst koji okružuje pogrešku. Primjeri takvih pogrešaka su sročnost (subjekt i predikat, član i imenica, atribut imenice itd.), problemi u glagolskom obliku (neispravno glagolsko vrijeme, modalitet itd.), podkategorizacija ili problemi u odabiru vrste riječi. U primjeru na slici 15. učenik je napisao za „grad“ *um cidade* kao da se radi o muškom rodu, dok je u portugalskom jeziku ispravan oblik *uma cidade* (ženski rod). Radi se o pogrešci u sročnosti koja je označena u pojavnici *um*. Treba obratiti pozornost i na polje *slemma* koje nije ispunjeno. Razlog za to je hijerarhijsko nasljeđivanje između razina, od niže (ortografska) prema višoj razini (leksička). Primjerice ako je polje *nform* prazno, sustav će pretpostaviti da je njegova vrijednost jednaka polju *form* (polje za učiteljske ispravke *fform* se ne nasljeđuje). To je još jedna od prednosti sustava anotacija koji pruža TEITOK, s obzirom da označivač mora unijeti isključivo nove (drugačije) podatke, a ne mora ispunjavati sva polja na svakoj razini.[146]



Slika 15: primjer ispravka gramatičke pogreške[147]

Treća se razina upotrebljava ako u učenikovom obliku riječi postoji leksički/semantički problem. Drugim riječima, učenikova riječ može biti gramatički i ortografski ispravna, ali nije prirodna u smislu da tu riječ izvorni govornik ne bi upotrijebio. Na slici 16. ispunjeno je samo polje *llema*, s obzirom da se samo ono razlikuje – *lpos* ima jednaku vrijednost kao *pos* te je stoga ostavljeno prazno.[148]

[146] del Rio, I.; Antunes, S.; Mendes, A.; Janssen, M. Towards error annotation in a learner corpus of Portuguese, 2016.
[147] Ibid.
[148] Ibid.



Slika 16.: prikaz ispravka leksičke pogreške[149]

Različite razine pogrešaka također omogućavaju različite vizualizacije teksta u kojima upisani ispravci zamjenjuje učenikove oblike riječi. Na taj način moguće je prikazati isti tekst s ispravcima na različitim razinama, od verzije bliže izvornom tekstu (samo ortografske ispravke) do verzije s najviše izmjena (ortografske, gramatičke i leksičke izmjene). S obzirom da navedeno označavanje funkcionira na razini pojavnice, takav sustav nije prikladan za pogreške koje obuhvaćaju više od jedne riječi, primjerice pogreške u redoslijedu riječi ili višerječni izrazi. Za takve slučajeve priprema se udaljena (*stand-off*) anotacija.[150]

Skup oznaka za pogreške

Kao sljedeći korak planirano je uvođenje kodova pogreški za svaku pogrešku označenu prema prethodno opisanom sustavu. Skup oznaka je u izradi, a za sada je definiran probni skup koji će se primijeniti na korpus radi testiranja učinkovitosti. Trenutačno se sastoji od 37 oznaka i strukturiran je u dvije razine informacija:

1) općenito lingvističko područje i
2) kategorija pogreške (po potrebi podkategorije).

Za prvu razinu upotrijebit će se prethodno spomenuta tri lingvistička područja. Uporabom istih općenitih lingvističkih područja za klasifikaciju pogrešaka omogućava se prijenos informacija između sustava s više razine i sustava kodova. U drugoj razini nalaze se uobičajene kategorije kao što su sročnost i kriva vrsta riječi. U svrhu izrade skupa oznaka proveden je eksperiment prilikom kojeg su pronađene pogreške i definirane kategorije potrebne za njihovo označavanje,

[149] del Rio, I.; Antunes, S.; Mendes, A.; Janssen, M. Towards error annotation in a learner corpus of Portuguese, 2016.
[150] Ibid.

a uključene su i neke za koje se očekuje da će se pojaviti u korpusu s obzirom na slične skupove oznaka razvijene za druge projekte. Oznake su organizirane tako da prvo slovo označava prvu razinu, dok sva naknadna slova označavaju drugu. Primjerice pogreške u sročnosti koje se odnose na rod imaju oznaku „GAG“, što označava *grammar* + *agreement* + *gender* (gramatika + sročnost + rod). Informacije o vrsti riječi nisu uključene jer se kod pogreške za pojavnicu dodaje u XML datoteku koja već sadržava te informacije.[151]

Kada je to moguće, informacije kodirane za svaku pojavnicu u korpusu upotrijebit će se za automatsko dodjeljivanje koda pogreške, tako da se usporede izvorni učenikov oblik s ispravcima (uz lemu i vrstu riječi) unesenima na razini anotacije pogrešaka. Prvo slovo u kodu dodjeljuje se automatski uzimajući u obzir razinu na kojoj je pogreška označena u sustavu s više razina (ortografska, gramatička ili leksička). Ostala slova u kodu dodijelit će se automatski po mogućnosti. Primjerice ispunjeno polje *nform* (ortografska razina) znači da se radi o ortografskoj pogrešci. To omogućava automatski klasifikaciju pogreške na lingvističkoj razini i dodjeljivanje prvog slova (u ovom slučaju S). Na sljedećoj razini može postojati kod za primjerice pogrešku u dijakritičkim znakovima (također S). Ako se nakon usporedbe učenikova oblika riječi i polja *nform* utvrdi da je razlika isključivo u dijakritičkim znakovima, moguće je automatski dodijeliti slovo za vrstu pogreške u kod (SS). Za većinu složenijih vrsta pogrešaka takva automatska klasifikacija neće biti moguća, ali će svejedno osobama koje rade na označavanju korpusa uštedjeti mnogo vremena.[152]

[151] del Rio, I.; Antunes, S.; Mendes, A.; Janssen, M. Towards error annotation in a learner corpus of Portuguese, 2016.
[152] Ibid.

6. Zaključak

Korpusi danas bez dvojbe zauzimaju važno mjesto kao alat u raznim lingvističkim istraživanjima. Interes za učenjem stranih jezika potaknuo je dodatnu specijalizaciju i razvoj učeničkih korpusa. Veći svjetski jezici njihovu su važnost brzo prepoznali te su u skladu s tim najbrojniji učenički korpusi engleskoga jezika. Objava korpusa *International Corpus of Learner English* otvorila je vrata razvoju učeničkih korpusa i na osnovu njega nastali su mnogi istraživački radovi i novi korpusi. Osim razvoja učeničkih korpusa za druge veće jezike poput francuskog, španjolskog i njemačkog, u zadnjih nekoliko godina došlo je do razvoja učeničkih korpusa i manjih, ali nama bližih slavenskih jezika poput češkog i slovenskog, a naposljetku i korpusa za sam hrvatski jezik. Međutim, da bi brojni korpusi koji nastaju u posljednje vrijeme mogli biti temelj za korisne lingvističke analize koje će doprinijeti razumijevanju međujezika i procesa učenja stranog jezika, potrebno je prije same izrade razmisliti o brojnim pitanjima. Neka od tih pitanja odnose se na same učenike (dob, spol, materinski jezik), dok se druga odnose na vrstu teksta koji se prikuplja i način prikupljanja i obrade. Česta je pogreška mnogih lingvista i drugih istraživača uključenih u izradu učeničkih korpusa preskakanje koraka određivanja bitnih varijabli u fazi projektiranja korpusa, što za posljedicu ima nedostatak formalnog postupka izrade. Upotrebljivost korpusa izravno ovisi o pažnji koja je posvećena određivanju varijabli i načinu prikupljanja građe. Iz tog razloga oportunistički prikupljeni korpusi teško mogu predstavljati temelj za neke relevantne lingvističke analize i usporedbe jezika učenika i izvornih govornika. Iako je iz svega navedenog vidljivo da je standardizacija postupka izrade i anotacije korpusa od presudne važnosti za daljnji razvoj u području, opće prihvaćeni standard prema kojem bi se izrađivali i označavali svi novi korpusi još uvijek ne postoji.

7. Literatura

- Bennett R., G. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. University of Michigan Press, 2010.

- Bratanić, M. Korpusna lingvistika na kraju 20. stoljeća i implikacije za suvremenu hrvatsku leksikografiju. // Filologija. 30 – 31 (1998), str. 171-177.

- Christodouloupoulos, C.; Steedman, M. A massively parallel corpus: the Bible in 100 languages. // Language Resources and Evaluation. 2014.

- Contrastive analysis. // Wikipedia: the free encyclopedia. 3.6.2017. Dostupno na: https://en.wikipedia.org/wiki/Contrastive_analysis (19.7.2017.).

- Corpus Encoding Standard. // Dostupno na: https://www.cs.vassar.edu/CES/CES1-0.html (28. 9. 2017.)

- Corpus Escrito de Espanol L2. // Dostupno na: http://www.uam.es/proyectosinv/woslac/collaborating.htm (19. 7. 2017.).

- Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2) // Dostupno na: http://teitok.iltec.pt/peapl2/index.php?action=home (28. 9. 2017.)

- del Rio, I.; Antunes, S.; Mendes, A.; Janssen, M. Towards error annotation in a learner corpus of Portuguese, 2016.

- Error analysis. // Wikipedia: the free encyclopedia. 3.5.2017. Dostupno na: https://en.wikipedia.org/wiki/Error_analysis_(linguistics) (19. 7. 2017.).

- Glaznieks, A.; Nicolas, L.; Stemle, E.; Abel, A.; Lyding, V. Establishing a Standardised Procedure for Building Learner Corpora. // Apples – Journal of Applied Language Studies. Vol. 8, 3, 2014. str. 5-20

- Granger, S. Computer learner corpus research: current status and future prospects. // Applied Corpus Linguistics: A Multidimensional Perspective / Ulla Connor, Thomas A. Upton. Amsterdam : Atlanta : Rodopi, 2004. str. 123-145.

- Granger, S. A Bird's-eye view of learner corpus research. // Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. / Sylviane Granger, Joseph Hung, Stephanie Petch-Tyson. Amsterdam/Philadelphia : John Benjamins, 2002. str. 3-33

- Hana, J.; Rosen, A.; Škodova, S.; Štindlova, B. Error-tagged Learner Corpus of Czech. // Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010. Uppsala, Švedska, 2010. str. 11–19

- Hunston, S. Corpus linguistics. // Encyclopedia of Language & Linguistics (Second Edition), 2006., str. 234-248.

- Ide, N. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. 1999.

- Janssen, M. TEITOK: Text-Faithful Annotated Corpora. // Proceedings of the Tenth International Conference on Language Resources and Evaluation, 2016.

- Klobučar Srbić, I. Obol korpusne lingvistike suvremenoj leksikografiji. // Studia lexicographica. 2 (3) (2008), str. 39-51.

- Kovář, V.; McCarthy, D. New Learner Corpus Functionality in the Sketch Engine. // Proceedings of the 2012 Asia Pacific Corpus Linguistics Conference, 2012.

- Kosem, I.; Kovar, V.; Baisa, V.; Kilgarriff, A. The Sketch Engine interface for a learner corpus annotated with errors and corrections. // Learner Corpus Research 2013, 2013.

- Kotani, K.; Yoshimi, T.; Nanjo, H.; Isahara, H. Corpus Materials for Constructing Learner Corpus Compiling Speaking, Writing, Listening, and Reading Data. // IJCLA vol. 3, br. 2, 2012. str. 77-92

- Learner corpora around the world. // Université catholique de Louvain. Dostupno na: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (19. 7. 2017.)

- Lozano, C.; Mendikoetxea, A. Learner corpora and Second Language Acquisition: The design and collection of CEDEL2. // Automatic Treatment and Analysis of Learner Corpus Data. / A. Díaz-Negrillo, N. Ballier i P. Thompson. Amsterdam: John Benjamins, 2013. str. 65-100.

- Mark, K.L. The Significance of Learner Corpus Data in Relation to the Problems of Language Teaching. // Bulletin of General Education 312, 1998. str. 77-90.

- McEnery, T.; Xiao, R.; Tono, Y. Corpus-Based Language Studies: an advanced resource book. New York, Routledge, 2006.

- Mendes, A.; Antunes, S.; Janssen, M.; Gonçalves, A. The COPLE2 Corpus: a Learner Corpus for Portuguese. 2016.

- Mikelić Preradović, N.; Berać, M.; Boras, D. Learner Corpus of Croatian as a Second and Foreign Language. // Multidisciplinary Approaches to Multilingualism. / Petar Lang, 2015.

- Miličević, M. Korpus srpskog kao stranog jezika (KSKS): Opis građe i plan izrade. // Srpski kao strani jezik u teoriji i praksi III / V. Krajišnik (ed.). Beograd: Filološki fakultet, 2016. str. 279-289

- Pejić, T. Računalni učenički korpusi i učenički korpus hrvatskog jezika. // Diplomski rad. Filozofski fakultet u Zagrebu, Odsjek za informacijske i komunikacijske znanosti. 2015.

- Second-language acquisition. // Wikipedia: the free encyclopedia. 26. 6. 2017. Dostupno na: https://en.wikipedia.org/wiki/Second-language_acquisition (19. 7. 2017.)

- Sinclair, J. Corpus, concordance, collocation. Oxford : Oxford University Press, 1991.

- Sketch Engine. // Wikipedia: the free encyclopedia. 21. 9. 2017. Dostupno na: https://en.wikipedia.org/wiki/Sketch_Engine (28. 9. 2017.)

- Sketch Engine. // Dostupno na: https://www.sketchengine.co.uk/ (28. 9. 2017.)

- Stritar, M. Slovene as a Foreign Language : The Pilot Learner Corpus Perspective. // Slovenski jezik – Slovene Linguistic Studies. 7 (2009), str. 135-152.

- Štrkalj Despot, K.; Möhrs, C. Pogled u e-leksikografiju. // Časopis Instituta za hrvatski jezik i jezikoslovlje. 41/2 (2015.). str. 329-353

- TEI By Example. // Dostupno na: http://teibyexample.org/TBE.htm (28. 9. 2017.)

- Teubert, W.; Čermakova, A. Corpus linguistics: a short introduction. London : New York : Continuum, 2007.

- Teubert, W. Language Resources: The Foundations of a Pan-European Information Society. // Trans-European Language Resources Infrastructure. Proceedings of the First European Seminar: „Language Resources for Language Technology“. (1995) str. 105-128

- Text Encoding Initiative. // Dostupno na: http://www.tei-c.org/index.xml (28. 9. 2017.)

- Tono, Y. Learner corpora: design, development and applications. 2003.

- Written Corpus of Learner English. // Dostupno na: http://web.uam.es/proyectosinv/woslac/Wricle/ (19. 7. 2017.)