SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET

ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI

AK. GOD. 2014. / 2015.

Tea Pejić

Računalni učenički korpusi i učenički korpus hrvatskog jezika

diplomski rad

Mentorica: dr. sc. Nives Mikelić Preradović, izv. prof.

Zagreb, rujan 2015.

Sadržaj

1. Uvod

U vremenu globalizacije, stalnih migracija stanovništva, globalne umreženosti i rapidnog tehnološkog napretka prisutna je rastuća potreba za jezičnim resursima, osobito za pomagalima za učenje stranih jezika. Novo doba donijelo je pogled na jezik iz novih, prije, primjerice, samo pola stoljeća nezamislivih perspektiva. „(...) We would be justified in calling CLC[1] research "a new research enterprise, a new way of thinking about *learner* language, which is challenging some of our most-deeply rooted ideas about *learner* language."[2] Moderna tehnologija omogućila je dostupnost kao i brzu obradu velikih količina podataka što će korpuse od malih zbirki tekstova pretvoriti u ogromne tekstualne arhive od više stotina milijuna riječi.

Od početaka razvoja korpusa 60-ih do 90-ih godina prošlog stoljeća kad dolazi do razvoja učeničkih korpusa pa sve do današnjih dana raste svijest o mnoštvu prednosti i mogućnostima napretka koje nude korpusi općenito, osobito učenički. Ovaj rad nastoji prikazati korpuse općenito, s naglaskom na učeničke korpuse i *Hrvatski učenički korpus*. Rad je ustrojen na način da kreće od općeg prema posebnom: Započinje definicijom računalne korpusne lingvistike kao područja u čijim okvirima je težište ovog rada, a to su računalni učenički korpusi. U prvom dijelu rada dan je opći pregled korpusa počevši od njihove definicije preko vrsta do povijesnog razvitka u svijetu i na našim prostorima, a zatim se prelazi na učeničke korpuse općenito te naposljetku na *Hrvatski učenički korpus*. U drugom dijelu rada dana je analiza istraživanja provedenog na dijelu spomenutog korpusa.

[1] Kratica CLC (computer learner corpus) odnosi se na računalne učeničke korpuse. O ovoj temi još će biti govora u nastavku.
[2] Granger 2004., str. 123

2. Korpusna lingvistika

2.1. Definicija

Suvremena korpusna lingvistika zasebna je grana lingvistike koja se bavi jezičnom analizom strojno izrađenih korpusa pisanoga ili govornoga jezika.[3] Sam pojam korpusne lingvistike nastao je prije tridesetak godina, iako su istraživanja u tom području počela još šezdesetih godina.[4] Ovu znanstvenu disciplinu odlikuje rad s „autentičnim jezičnim podacima“[5], tj. prilikom dolaska do novih spoznaja bilo o jeziku općenito bilo o određenim jezicima, temelj su kvantitativni ili kvalitativni podaci dobiveni lingvističkim analizama tekstualnih ili korpusa govornog jezika[6] stvorenih iz neovisnih izvora poput književnosti, stručne literature, novinskih članaka, govora, spontanih razgovora[7] itd. Prema tome, spoznaje u korpusnoj lingvistici ne temelje se na jeziku kao apstraktnoj pojavi, nego na stvarnoj jezičnoj praksi, tj. na onom jeziku koji je doista u uporabi. Cilj korpusne lingvistike jest na temelju korpusa dokazati odnosno opovrgnuti postojeće hipoteze te stvoriti nove. U skladu s tim Elena Tognini-Bonelli[8] razlikuje 2 pristupa:

- Korpusno upravljan pristup (*corpus-driven approach)* – podrazumijeva lingvističku analizu uz pomoć korpusa odnosno hipoteze se ekstrahiraju isključivo iz tekstualnih materijala korpusa.
- Korpusno utemeljen pristup (*corpus-based approach)* – lingvistička analiza nastaje na temelju korpusa što znači da korpus služi i za istraživanje i opis jezičnih zakonitosti, ali i za provjeru postojećih hipoteza i teorija[9].

Iz temeljne pretpostavke od koje polazi korpusna lingvistika, a to je stvaranje novih spoznaja na temelju izvornih jezičnih podataka, proizlazi nekoliko metodičkih problema. Jezik kao predmet istraživanja beskonačan je, stalno se mijenja i nemoguće je „zalediti“ ga i da on pritom ostane reprezentativan kao jezična stvarnost. Upravo to događa se pri stvaranju korpusa. Ovaj se problem nastoji riješiti stvaranjem korpusa sve većeg opsega pa tako korpus, kako bi analize provedene na njemu bile relevantne, mora sadržavati više od milijun

[3] Klobučar Srbić 2008., str. 39
[4] Bratanić 1992., str. 145
[5] Berman, S. Korpus und Korpuslinguistik. // Ruhr-Universität Bochum. (4.7.2015.).
[6] Korpuslinguistik. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[7] Berman, S. Korpus und Korpuslinguistik. // Ruhr-Universität Bochum. (4.7.2015.).
[8] Teubert; Čermakova 2007., str. 57
[9] Bopp, S. Einführung in die Korpuslinguistik mit DeReKo und COSMAS II. // Philologische und historische Fakultät der Universität Augsburg. (4.7.2015.)., str. 6

pojavnica.[10] Koliko je neki korpus reprezentativan može se ocijeniti i ako se u obzir uzme njegova svrha. To se nastoji postići pomnim odabirom tekstova unutar određenog područja prilikom sastavljanja korpusa.[11] Nadalje, korpus kao takav uključuje i promišljeno oblikovane pisane tekstove, ali i spontane konverzacije koje dovode do odstupanja od jezičnog standarda te je stoga u nekim slučajevima kad imamo rezultate koji pokazuju da se neki jezični fenomen pojavljuje u korpusu teško utvrditi je li riječ o stvarnoj jezičnoj praksi ili o iznimkama. Drugi problem koji je neposredno vezan uz ovaj povezan je s gore navedenom činjenicom o beskonačnosti i konstantnoj promjenjivosti leksika pa se neke pojave u jeziku ne mogu dokazati ni istraživati pomoću analize korpusa, jer nijedan korpus nije toliko opsežan da bi obuhvatio cjelokupan leksik nekog jezika. A to opet ne mora značiti da takve pojave ne postoje u jeziku ni da su pogrešne u gramatičkom smislu.

Budući da se temelji na uporabi prirodnih jezika, korpusna se lingvistika smatra induktivnom metodom te se ubraja u empirističke metode – uz pomoć što je moguće više konkretnih pojedinačnih primjera stvara se opći zaključak.[12] Time je, kao i kognitivna lingvistika, u suprotnosti s vladajućom paradigmom generativne gramatike Noama Chomskog koja previše pozornosti ne pridaje značenju, nego je fokusirana na gramatiku[13] odnosno jezičnu kompetenciju idealnog izvornog govornika[14]. Chomsky odbacuje upotrebu korpusa[15] i bilo kakvog jezičnog empirijskog istraživanja koje nije utemeljeno na kompetenciji izvornog govornika[16]. Smatra da iz korpusa saznajemo samo sadašnje stanje nekog jezika, no ne i kakav će biti u budućnosti.[17] Bitnim svojstvom gramatike smatra činjenicu omogućuje li ona stvaranje beskonačnog skupa gramatičkih rečenica.[18] Chomsky drži da se jezični procesi odvijaju odvojeno od ostalih mentalnih aktivnosti, dok, primjerice, kognitivna lingvistika jezične procese smatra sastavnim dijelom cjelokupne mentalne strukture. Stoga je, kad govorimo o jeziku, nužno razmotriti uporabu jezika, jer iz same uporabe proizlazi svaki pristup jeziku uz pomoć korpusa.[19]

[10] Klobučar Srbić 2008., str. 40
[11] Bratanić 1992., str. 154
[12] Korpuslinguistik. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[13] Teubert; Čermakova 2007., str. 47
[14] Korpuslinguistik. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[15] Generativna gramatika. // Hrvatska enciklopedija. (4.7.2015.).
[16] Bratanić 1992., str. 146
[17] Teubert; Čermakova 2007., str. 47
[18] Generativna gramatika. // Hrvatska enciklopedija. (4.7.2015.).
[19] Klobučar Srbić 2008., str. 39

2.2. Lingvistička metoda ili samostalna disciplina

Još uvijek nije konačno definirano je li korpusna lingvistika tek „puka metoda“[20] opće ili primijenjene lingvistike[21] koja je uvođenjem uporabe računala lingvistima uvelike olakšala posao ili je pak riječ o „novoj lingvistici“, tj. novonastaloj lingvističkoj disciplini. „Je li ona tek jedna usamljena struja ili pak rijeka sve snažnijeg toka u kojoj se već prilično jasno razabire autohton pristup istraživanju jezika?“[22] Postoje argumenti koji podupiru oba stajališta. Tako korpusnu lingvistiku možemo smatrati lingvističkom metodom, jer se s jedne strane mnoge grane lingvistike služe empirijskom analizom korpusa, a s druge strane možemo konstatirati da korpusna lingvistika nema točno određen predmet proučavanja pa je se u skladu s tim teško može smatrati zasebnom disciplinom. No, svojim predmetom spoznaje smatra stvarnu jezičnu uporabu i na taj se način ograđuje od drugih lingvističkih disciplina koje se bave jezičnom sposobnošću čovjeka ili općim jezičnim strukturama. Unatoč navedenim dvojbenostima, korpusna se lingvistika učvrstila kao samostalna znanstvena disciplina što svakako potvrđuje izlaženje tematskih stručnih časopisa te osnutak dviju katedri na sveučilištima u Birminghamu i u Berlinu.[23]

2.3. Ciljevi, metodologija i alati korpusne lingvistike

Na temelju izvornih tekstova[24] sakupljenih u korpus korpusna lingvistika nastoji pomnije istražiti sam jezik i njegovu stvarnu uporabu. Pritom se služi suvremenim metodama kako bi se pojednostavio postupak istraživanja te omogućio jednostavniji pristup jezičnim podacima.[25] Glavno pomagalo, možemo reći i temelj razvoja korpusne lingvistike, jest računalo koje služi za „usustavljivanje i pretraživanje građe“ te za „provjeru istraživačkih hipoteza“[26]. Neki od najpoznatijih sustava za rad s korpusima su OCP ‑ *Oxford Concordance Program, XML (Extensible Markup Language, XCES (Corpus Encoding Standard for XML)* i Unicode.[27]

[20] Bratanić 1992., str. 145
[21] Korpuslinguistik. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[22] Bratanić 1992., str. 145
[23] Korpuslinguistik. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[24] Bratanić 1992., str.151
[25] Klobučar Srbić 2008., str. 41
[26] Tadić 1996., str. 604
[27] Klobučar Srbić 2008., str. 41

3. Korpusi u općenitom smislu

3.1. Definicija

Korpusna lingvistika proučava značenje riječi kao osnovne jezične jedinice.[28] Jezična zajednica definira upotrebu i značenje riječi. No, pritom ne dolazi uvijek do 'sporazuma' – unutar zajednice može doći do neslaganja oko značenja ili konotacija nekih riječi.[29] Jezične se zajednice smatraju socijalnim konstruktima, a mogu se definirati na temelju iskustva, jer mnoge od njih imaju specifična obilježja. Tako, primjerice, jezik kojim se govorilo u bivšoj Jugoslaviji tzv. srpskohrvatski danas „ne postoji“, tj. govorimo o srpskom i hrvatskom kao o dva različita jezika koji su uvijek postojali kao takvi. Stoga nije moguće definirati jedinstvena pravila o tome što je jezična zajednica, a što nije.[30] Sudionici diskursa odlučuju o tome što uvesti u diskurs, tj. što oni smatraju činjenicama.[31] U diskursu možemo sudjelovati u ulozi govornika i u ulozi slušatelja. Ako kao član jezične zajednice određenu riječ ili frazu ne smatramo dovoljno dobrom da bi opisala ono što njome želimo izreći, možemo, ili kreirati novu riječ i pokušati je uvesti u uporabu u jezičnoj zajednici, ili postojeću upotrijebiti u novom kontekstu.[32] No, da bi nas se razumjelo kao govornike, moramo se držati pravila odnosno govoriti u skladu s očekivanjima govornika, tj. u skladu s dosadašnjom jezičnom praksom.[33] Diskurs čine svi tekstovi – tiskani ili u rukopisu, izgovoreni (većina ih je izgubljeno tijekom povijesti, jer ih se nije moglo snimiti – tek u novije vrijeme pojavljuje se govoreni tekst).[34] Korpus kao takav trebao bi predstavljati diskurs ili neki njegov dio. Tako, primjerice, na temelju Brown korpusa možemo vidjeti stanje engleskog jezika iz 1961. godine.[35] Jezik odnosno diskurs podrazumijeva sve verbalne interakcije koje su se odvile ili se odvijaju u jezičnoj zajednici. U jezične interakcije neke zajednice ubrajaju se idiolekti, dijalekti, sociolekti, regionalne varijacije, suvremeni i jezik iz minulih vremena, sleng i žargon.[36]

Neki leksikografi tvrde da stanje u korpusu nije prikaz stvarnog stanja u jeziku, nego slučajna i proizvoljna 'zbirka pojavnica', no profesor korpusne lingvistike sa Sveučilišta u

[28] Teubert; Čermakova 2007., str. 50
[29] Teubert; Čermakova 2007., str 48
[30] Teubert; Čermakova 2007., str 61-62
[31] Teubert; Čermakova 2007., str 41
[32] Teubert; Čermakova 2007., str. 78
[33] Teubert; Čermakova 2007., str. 81
[34] Teubert; Čermakova 2007., str 46
[35] Teubert; Čermakova 2007., str. 59
[36] Teubert; Čermakova 2007., str. 61

Birminghamu Wolfgang Teubert smatra da se korpus može promatrati kao jezična stvarnost pretočena u tekst. [37]

Korpus je moguće definirati na nekoliko načina. Općenito govoreći, može se reći da je korpus zbirka[38] pisanih ili u pismu zabilježenih usmenih[39] jezičnih odsječaka[40] prirodnoga jezika sastavljena po određenim kriterijima[41] kako bi činila jezični uzorak. Treba napomenuti da su i sami tekstovi koji čine korpus skupljeni prema određenim znanstvenim[42] kriterijima[43], primjerice, određeni broj i određena vrsta tekstova čini sastav nekog korpusa[44], iz čega proizlazi da se svaka zbirka tekstova ne može smatrati korpusom.[45] Prema definiciji Johna Sinclaira, korpus je zbirka tekstova nekog jezika nastala na prirodan način, kojom se opisuje raznolikost određenog jezika[46].

3.2. Vrste

Postoji više različitih podjela korpusa s obzirom na formalne i sadržajne[47] kriterije poput funkcionalnosti, broja jezika, medija, veličine, anotacije, persistentnosti, dostupnosti[48] itd. Osnovna podjela korpusa jest ona na računalne korpuse i korpuse u papirnatom obliku. Danas su uglavnom računalni korpusi u uporabi, no korpusi u papirnatom obliku ranije su imali vrlo važnu ulogu, primjerice, prilikom izrade rječnika.[49]

Nadalje, s obzirom na dio jezika koji obuhvaćaju, korpuse možemo podijeliti na referentne i specijalne. Referentni korpusi nastoje obuhvatiti cjelokupan jezik u nekom vremenskom razdoblju[50] pa stoga takvi korpusi sadrže nekoliko milijuna ili čak nekoliko milijardi pojavnica.[51] Referentni korpusi sastoje se od onih tekstova za koje su se članovi zajednice složili da su reprezentativni za standardni jezik. Ranije se standardni jezik temeljio na jeziku „dobre“ književnosti. Danas bi se standardni jezik moglo definirati kao skup svih tekstova koje pripadnici obrazovanog srednjeg sloja pročitaju odnosno čuju u jednoj godini. Mogao bi uključivati ozbiljne novine, tabloide, časopise, knjige te svojevrsnu mješavinu računa,

[37] Teubert; Čermakova 2007., str 41
[38] Klobučar Srbić 2008., str. 39
[39] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[40] Klobučar Srbić 2008., str. 39
[41] Sinclair 1991., str. 17
[42] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[43] Klobučar Srbić 2008., str. 39
[44] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[45] Klobučar Srbić 2008., str. 39
[46] Teubert; Čermakova 2007., str. 62
[47] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[48] Lemnitzer; Zinsmeister 2006., str. 103
[49] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[50] Lemnitzer; Zinsmeister 2006., str. 106
[51] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).

brošura, uputa za uporabu te govorni tekst koji čuju u razgovorima, na društvenim događanjima ili u medijima. Referentni korpusi pojedinih jezika mogu se uspoređivati ako su približno iste veličine te ako su žanrovi tekstova koje sadrže slični.[52] Budući da zbog svojeg opsega i uravnoteženosti[53] predstavljaju jezik u cjelini, koriste se u najrazličitije svrhe. Omogućuju uvid u jezičnu raznolikost na svim razinama jezika[54], sadrže vokabular standardnog jezika pa su stoga najbolji izvor informacija o značenju[55], mogu se upotrebljavati i u semantičkoj analizi kako bi se izdvojila različita značenja nekoga leksema[56]. Nadalje, ako su referentni korpusi dovoljno opsežni, u njima nalazimo fraze sa svim njihovim značenjima, koje nam o značenju govore mnogo više od riječi zasebno (npr. Britanski nacionalni korpus). Upotrebljavaju se i prilikom istraživanja specijalnih korpusa[57], tj. kao kontrolni korpus[58] u okviru istraživanja nekog specifičnog fenomena unutar standardnog jezika.[59] Valja napomenuti da se, osim u raznim disciplinama same lingvistike, korpusi uvelike upotrebljavaju u istraživanjima i u drugim znanostima poput, primjerice, povijesti ili književnosti.[60]

Najpoznatiji referentni odnosno nacionalni korpusi, osim Britanskog nacionalnog korpusa, su *Njemački nacionalni korpus* koji se izrađuje na Institutu za njemački jezik u Mannheimu[61], *Češki nacionalni korpus*, *Ruski nacionalni korpus* itd.[62]

Za razliku od referentnih korpusa, na specijalnim se korpusima proučavaju specifični fenomeni u jeziku. Sinclair, primjerice, u tu kategoriju ubraja jezik djece, starijih osoba, govornika dijalekata takoreći udaljenih od standarda te jezike različitih struka odnosno područja.[63]

Profesor Teubert navodi oportunistički korpus kao još jednu vrstu korpusa, koja je suprotna referentnom – ne teži reprezentativnosti niti mu je cilj prikazati jezični diskurs. Temelji se na pretpostavci da je svaki korpus neuravnotežen te stoga problem reprezentativnosti i

[52] Teubert; Čermakova 2007., str. 66
[53] Klobučar Srbić 2008., str. 48
[54] Klobučar Srbić 2008., str. 40
[55] Teubert; Čermakova 2007., str. 67
[56] Klobučar Srbić 2008., str. 48
[57] Teubert; Čermakova 2007., str. 68
[58] Lemnitzer; Zinsmeister 2006., str. 106
[59] Teubert; Čermakova 2007., str. 68
[60] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[61] Korpuslinguistik. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[62] Klobučar Srbić 2008., str. 44
[63] Granger 1998.

uravnoteženosti možemo promatrati uvijek iz drugog kuta. Što je oportunistički korpus opsežniji, to bolje.[64]

S obzirom na persistentnost, korpuse dijelimo na statične i monitor korpuse[65]. Statični korpusi podrazumijevaju određenu zbirku tekstova koja je skupljena u određenom vremenskom razdoblju te je spremna za uporabu u određene svrhe. Većina korpusa funkcionira kao statičan korpus, no takvi korpusi ne moraju zauvijek ostati u tom stanju. Neki od njih mogu se nadopunjavati u određenim vremenskim razmacima pa za takve korpuse možemo reći da su zapravo dinamični.[66] Primjer statičnog korpusa su korpusi koji sadrže djela preminulih autora ili tekstove napisane na nekom od mrtvih jezika.[67] Pod monitor korpusom[68] podrazumijevamo korpus koji redovito prati promjene u jeziku, osobito leksičke poput, primjerice, promjena učestalosti pojavljivanja riječi, složenica, sintagmi, fraza što je obično znak da je došlo do promjene u značenju. Bilježi i promjene poput pojave novih riječi, novih značenjskih jedinica te promjena konteksta u kojima se pojavljuju riječi ili druge jezične jedinice.[69] Primjer monitor korpusa bio bi korpus sastavljen od članaka nekih (još uvijek aktivnih) dnevnih novina[70], jer su kao takve najbolji i najbrži pokazatelji promjena u jeziku[71].

Što se tiče anotacije korpusa, postoje obilježeni i sirovi korpusi. Sirovi korpusi sadrže isključivo jezične podatke, dok obilježeni korpusi, osim jezičnih podataka, sadrže i dodatne podatke – metapodatke.[72] Korpusi mogu biti obilježeni na više lingvističkih razina kao što su morfosintaksa, sintaksa, semantika, pragmatika[73] i sl., ali mogu sadržavati i druge metapodatke poput vremena nastanka nekog teksta, imena autora, metapodatke o izradi korpusa[74] itd.

Kad govorimo o broju jezika u korpusu, razlikujemo jednojezične, dvojezične i višejezične korpuse.[75] Jednojezični korpusi, kako sam naziv govori, sadrže tekstove samo na jednom jeziku[76] pri čemu treba voditi računa o različitim varijantama dotičnog jezika ili o dijalektima. Kad je riječ o dvojezičnim ili višejezičnim korpusima, možemo dalje razlikovati paralelne i

[64] Teubert; Čermakova 2007., str. 70-71
[65] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[66] Lemnitzer; Zinsmeister 2006., str. 105-106
[67] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[68] Sinclair 1991., str. 25-26
[69] Teubert; Čermakova 2007., str. 71-72
[70] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[71] Teubert; Čermakova 2007., str. 72
[72] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[73] Lemnitzer; Zinsmeister 2006., str. 105
[74] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[75] Lemnitzer; Zinsmeister 2006., str. 103
[76] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).

usporedive korpuse.[77] Paralelni korpus sadrži tekstove na jednom jeziku i njihove prijevode na drugi jezik (ili više njih). Ponekad takav korpus sadrži samo prijevode tekstova na različite jezike, no ne i izvorni tekst. Za paralelne bi se korpuse moglo reći da predstavljaju repozitorij rada prevoditelja. Čak ni najopsežniji dvojezični rječnici u sebi ne mogu sadržavati toliko prijevodnih ekvivalenata koliko ih možemo naći u paralelnim korpusima, a koji su u stvarnoj jezičnoj uporabi[78]. Dvojezični rječnici uglavnom nude prijevod pojedinih riječi, sadrže dakako i višečlane jezične jedinice, no ne mogu ponuditi prijevodni ekvivalent unutar određenog, specifičnog konteksta koji je zapravo ključan čimbenik prilikom prevođenja. Ova činjenica onima koji određeni jezik ne poznaju dobro onemogućuje prevođenje samo na temelju rječnika. Samo prevođenje koristeći se jednim ili drugim pomagalom posve je različito: Koristeći se rječnikom, prevoditelj traži ekvivalent za riječ, dok prevoditelj koji se koristi paralelnim korpusom ima pristup različitim kontekstima u kojima se određena jezična jedinica može pojaviti te imajući to pred sobom, može bolje procijeniti koji bi ekvivalent odgovarao kontekstu teksta koji prevodi.[79] Kako bi bili što jednostavniji za uporabu, paralelni korpusi se poravnavaju[80] što znači da se izvornik i njegov prijevod poravnaju jedan s drugim, tako da rečenica izvornika i dotična rečenica prijevoda budu jedna nasuprot drugoj[81]. Usporedni korpusi su korpusi koji sadrže tekstove iz srodnih područja na različitim jezicima, a koji nisu međusobni prijevodi.[82]

Naposljetku, i internet se može smatrati nekom vrstom korpusa.[83] Internet, doduše, jest zbirka tekstova, no ipak se ne može smatrati korpusom u lingvističkom smislu[84], jer ga se ne može smatrati reprezentativnim, iako još uvijek nije dokazano postižu li se bolji rezultati istraživanjem velikog korpusa sastavljenog od općenitih tekstova ili korpusa manjeg opsega, ali sačinjenog od relevantnih tekstova iz ciljanog područja[85]. Unatoč tomu, mnogi lingvisti služe se internetom kao virtualnim korpusom koji sadrži više tekstova od bilo koje knjižnice na svijetu. Omogućuje, primjerice, provjeru postoje li određene riječi, u kojim kontekstima i izvorima se pojavljuju. Nedostaci interneta u tom smislu jesu njegova nestalnost – tekstovi koje danas čitamo, možda sutra neće biti dostupni ili će se pojaviti novi, nesistematičnost,

[77] Lemnitzer; Zinsmeister 2006., str. 103-104
[78] Klobučar Srbić 2008., str. 48
[79] Teubert; Čermakova 2007., str. 73-76
[80] Lemnitzer; Zinsmeister 2006., str. 104
[81] Teubert; Čermakova 2007., str. 73-76
[82] Lemnitzer; Zinsmeister 2006., str. 104
[83] Teubert; Čermakova 2007., str. 76
[84] Textkorpus. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[85] Kilgarriff, A.; Grefenstette, G. 2003., str. 9

nepreglednost, dijelom i nepouzdanost, ali i veličina[86]. S jedne strane omogućuje brzo i jeftino objavljivanje različitih tekstova što s druge strane lako može prerasti u njegov nedostatak, jer velik broj tekstova objavljenih na internetu prate gramatičke i pravopisne pogreške kao i neprovjereni autori odnosno informacije.[87]

Stoga se preporučuje prethodno preuzimanje svih tekstova s interneta kako bi se na taj način oformio korpus.[88] No, internet nudi i mnoge prednosti za lingvistička istraživanja, primjerice, sadrži ogroman broj tekstova različitih vrsta i veličina, pisanih na različitim jezicima, podrazumijeva pretraživanje u elektroničkom obliku što ubrzava proces istraživanja, lako je dostupan i stalno u porastu. Stoga se smatra da će se u budućnosti mnogo više koristiti u znanstvenoistraživačke svrhe.[89] Također, očekuje se da će u budućnosti pretraživači poput, primjerice, *Googlea* omogućiti pretraživanje prilagođeno potrebama lingvističkih istraživanja koje će uključivati veći broj rezultata, prikaz šireg konteksta prilikom pretraživanja, davati preciznije statistike i relevantnije rezultate s obzirom na zadane kriterije pretraživanja, omogućavati pretragu prema specifičnim lingvističkim kriterijima poput vrsta riječi[90] i sl.

3.3. Povijest

3.3.1. Korpusi u svijetu

Korpusna lingvistika relativno je nov pristup jeziku nastao 60-ih godina prošlog stoljeća. Stanje u to vrijeme zahtijevalo je promjene – jezična svojstva bila su nedovoljno opisana, tj. bili su potrebni stvarni podaci o jeziku.[91] Zbog središnje uloge engleskog jezika u svijetu kao i njegove široke rasprostranjenosti, razvoj korpusne lingvistike započinje na području SAD-a. Shodno tomu i najveći broj radova iz ovog područja nastao je na engleskom jeziku.[92]

Prvi projekt prikupljanja jezičnih podataka za empirijsko gramatičko[93] istraživanje većih razmjera u engleskom govornom području bio je *Randolph Quirk's Survey of English Usage* koji još nije bio računalno obrađen[94]. Iz tog je projekta nastala gramatika engleskog standardnog jezika koja će biti u uporabi desetljećima.[95]

[86] Bickel 2006., str. 75-76
[87] Kilgarriff, A.; Grefenstette, G. 2003., str. 9
[88] Teubert; Čermakova 2007., str. 76-77
[89] Bickel 2006., str. 81-82
[90] Kilgarriff, A.; Grefenstette, G. 2003., str. 12
[91] Teubert; Čermakova 2007., str. 50
[92] Sinclair 1991., str. 2
[93] Teubert; Čermakova 2007., str. 51
[94] Bratanić 1992., str. 147
[95] Teubert; Čermakova 2007., str. 51

Osnivačima moderne računalne korpusne lingvistike smatraju se Henry Kučera i Nelson Francis. Na temelju njihova rada „Computational Analysis of Present-Day American English“[96] izrađen je tzv. Brown korpus[97], prvi računalno obrađen jezični korpus[98]. Izrađen je za engleski jezik[99] s ciljem istraživanja gramatike i vokabulara na Sveučilištu u Brownu, Sjedinjene Države te sadrži milijun pojavnica.[100]

No, korpusi prve generacije mogu se koristiti samo za fonološke, morfološke i sintaktičke analize, dok istraživanja na većim korpusima – prema Johnu Sinclairu – daju preciznije rezultate.[101] S napretkom tehnologije i korpusna lingvistika brzo napreduje pa već 80-ih i 90-ih godina nastaju tzv. korpusi druge generacije sastavljeni od nekoliko milijuna pojavnica.[102] Glavni problem vezan uz korpuse iz ovog razdoblja jest pitanje standardizacije – stvaranja zajedničkih pravila prema kojima bi se izgrađivali odnosno uređivali korpusi (kako će se kodirati, trebaju li se eksterne informacije dodavati u obliku bilježaka ili označavanjem, mogu li se stvorena pravila primjenjivati na sve jezike).[103] Tako 80-ih godina prošlog stoljeća John Sinclair u suradnji s izdavačkom kućom Collins Cobuild izrađuje korpus *COBUILD*[104] *(Collins Birmingham University International Language Database)* na temelju kojeg će nastati jednojezični rječnik engleskoga jezika.[105] Objavljen je 1987. i prvi je rječnik nastao isključivo na temelju korpusa[106]. Korpus obuhvaća za to vrijeme veliki broj riječi – 18,3 milijuna, no riječi koje se rjeđe upotrebljavaju ne mogu se naći u rječniku, jer se ne pojavljuju u korpusu. Značenja jezičnih jedinica poredana su po frekvenciji u korpusu[107], no raspon značenja pojedinih riječi nije definiran na temelju korpusa, nego su se leksikografi poslužili korpusom kako bi potvrdili svoja stajališta te za izradu primjera.[108] Isto tako, ovim se korpusom nastojala prikazati jezična uporaba ne samo u preskriptivnom (jezična uporaba definirana jezičnim pravilima), nego i u deskriptivnom smislu (stvarna jezična uporaba).[109]

[96] Korpuslinguistik. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[97] Sinclair 1991., str. 23
[98] Klobučar Srbić 2008., str. 44
[99] Korpuslinguistik. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[100] Teubert; Čermakova 2007., str. 51-52
[101] Bratanić 1992., str. 155
[102] Klobučar Srbić 2008., str. 44
[103] Teubert; Čermakova 2007., str. 54
[104] Sinclair 1991., str. 2-3
[105] Klobučar Srbić 2008., str. 44
[106] Teubert; Čermakova 2007., str 57
[107] Klobučar Srbić 2008., str. 44
[108] Teubert; Čermakova 2007., str 57
[109] Korpuslinguistik. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).

Glede obrade korpusa u to vrijeme općenito može se reći da raste zanimanje za semantičke osobitosti korpusa.[110] Najpoznatiji korpus druge generacije jest British National Corpus sastavljen od 100 milijuna pojavnica i izrađen u nekomercijalne svrhe.[111] British National Corpus je i danas reprezentativan za engleski jezik te služi kao uzor ostalim jezicima pri izradi nacionalnih korpusa.

Predviđa se da će korpusi treće generacije biti sastavljeni od stotine milijuna riječi i da će se izrađivati u komercijalne svrhe. S vremenom će korpusi kakve danas poznajemo prijeći u goleme arhive tekstova[112]. Primjerice, u ovu se skupinu ubraja spomenuti projekt COBUILD u svom modernoj inačici koja već danas sadrži više od 650 milijuna pojavnica.[113] Što se tiče budućnosti korpusne lingvistike općenito, očekuje se da će se i dalje razvijati korpusi te računalni programi namijenjeni njihovom istraživanju, no poseban napredak očekuje se u istraživanju govornog jezika te u višejezičnim istraživanjima. S razvojem tehnologije proširit će se i spektar mogućnosti koje nudi korpus pa će porasti i njegova primjena i izvan okvira same lingvistike.[114]

3.3.2. Korpusi u Hrvatskoj

3.3.2.1. Od početaka do 1990. godine

Općenito govoreći, prvi korpus na hrvatskom jeziku nastao je u okviru disertacije *Raznolikost rječnika. Struktura govora* psihologa Ivana Furlana. No, prvi hrvatski korpus izrađen u lingvističke svrhe te računalno obrađen nastao je na temelju djela *Osman* književnika Ivana Gundulića iz razdoblja baroka, a izradio ga je za svog boravka u SAD-u Željko Bujas 1967. koji je nakon *Osmana* izradio korpuse za još nekoliko djela iz starije i novije hrvatske književnosti. Po povratku iz SAD-a Bujas nastavlja u istom smjeru razvoja kad zajedno s Rudolfom Filipovićem na Filozofskom fakultetu u Zagrebu pokreće projekt pod nazivom *Yugoslav Serbo-Croatian – English Contrastive Project*. Spomenuti Brown korpus poslužio je kao temelj projekta, tj. sadržaj Brown korpusa prevodio se na tri standardne varijante

[110] Bratanić 1992., str. 152
[111] Korpuslinguistik. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).
[112] Klobučar Srbić 2008., str. 44
[113] The Collins Corpus. // Collins. (4.7.2015.).
[114] Bratanić 1992., str. 157

hrvatskog ili srpskog[115], kako se to tada zvalo[116]. Višejezičnost je omogućila stvaranje paralelnih korpusa koji će pružiti dodatne mogućnosti istraživanja.

Godine 1968. također u Zavodu za lingvistiku pod vodstvom Milana Moguša dolazi do pokretanja projekta *Jezik Marka Marulića*. Dvije godine kasnije projekt je proširen i otada nastavlja s radom pod nazivom *Kompjutorska analiza tekstova stare hrvatske književnosti*[117]. Ovim se projektom nastojalo obuhvatiti prije svega djela starijih hrvatskih pisaca poput Petra Zoranića, Hanibala Lucića, Marina Držića, Tituša Brezovačkog itd. Tako nastali korpus omogućio je mnoga istraživanja na svim jezičnim razinama, no isto tako javila se potreba za izradom korpusa suvremenog hrvatskog jezika koja je dovela do pokretanja projekta pod nazivom *Korpus suvremenog hrvatskog književnog jezika* 1976. godine. Cilj projekta bio je stvoriti korpus koji će obuhvaćati milijun pojavnica.[118] Iako korpus obuhvaća razdoblje od 1935. do 1978., što znači da je građa za njegovu izradu bila dostupna u potpunosti, završen je tek 1996. godine. Njegova izrada uključivala je abecedne i frekvencijske rječnike pojavnica, konkordancija i strojno potpomognute lematizacije.[119] Sastoji se od pet potkorpusa s obzirom na tekstne vrste koje oni sadrže: drama, novine, proza, stihovi, udžbenici. Ovaj korpus omogućuje jezična istraživanja na različitim jezičnim razinama, a posebno treba izdvojiti mogućnosti istraživanja koje nudi na leksičkoj razini, jer omogućuju usporedbu s leksikom otprije 1990. godine.

Do 1991. godine hrvatska je korpusna lingvistika išla ukorak sa svjetskim trendovima. Posebice Zavod za lingvistiku Filozofskog fakulteta u Zagrebu kao središnja institucija za istraživanja ove vrste ostvarivao je značajnu suradnju s važnim lingvističkim centrima u Europi poput Birminghama ili Mannheima. No, zbog ratnih zbivanja dolazi do stagnacije u razvoju.

3.3.2.2. HNK

Po završetku Domovinskog rata u okviru Zavoda za lingvistiku ponovno se pokreću različiti projekti, a tu svakako treba izdvojiti projekt pod nazivom *Računalna obradba hrvatskoga jezika* unutar kojeg je 1996. godine pod vodstvom Marka Tadića pokrenuta izrada velikog korpusa hrvatskog jezika koji po uzoru na slične projekte u svijetu dobiva naziv *Hrvatski nacionalni korpus* odnosno *HNK*.[120] Sam naziv nastao je po uzoru na nazive postojećih

[115] Tadić 1997., str. 388
[116] Srpskohrvatski jezik. // Wikipedija : slobodna enciklopedija. (4.7.2015.).
[117] Bratanić 1992., str. 149
[118] Tadić 1997., str. 389
[119] Tadić 1992., str. 174-175
[120] Tadić 1997., str. 390-392

reprezentativnih korpusa pojedinih jezika (npr. *English National Corpus*). Čine ga tzv. *30M*, korpus suvremenog hrvatskog jezika za koji je prilikom pokretanja bilo predviđeno da će dostići opseg od 30 milijuna pojavnica[121] i *HETA* odnosno *Hrvatski elektronski tekstovni arhiv*. Ovakva podjela na korpus i zbirku tekstova proizlazi iz potrebe da se postavi određena vremenska granica odnosno zbog zahtjeva reprezentativnosti[122]. Tako je kao „granica suvremenosti" uzeta 1990. godina zbog presudnih političkih događaja koji će hrvatskom jeziku omogućiti slobodniji razvoj u godinama koje su uslijedile.[123] U skladu s tim, korpus 30M sadrži tekstove nastale nakon 1990. godine, dok HETA sadrži tekstove koji su stariji od 1990. godine i tekstove koji ne zadovoljavaju kriterij reprezentativnosti[124] odnosno koji bi narušili uravnoteženost strukture 30M korpusa[125].

Što se tiče opsega, on ovisi prvenstveno o namjeni korpusa.[126] Budući da je HNK zamišljen kao referentni korpus, podrazumijeva se da je njegova namjena višestruka odnosno da omogućuje jezična istraživanja na svim razinama pa stoga mora imati određenu veličinu iz koje će proizaći njegova raznolikost. Umjesto očekivanih 30 milijuna pojavnica[127], Hrvatski nacionalni korpus danas sadrži 216,8 milijuna pojavnica[128].

Vratimo se definiciji korpusa kao zbirke jezičnih odsječaka izrađene prema preciznim lingvističkim parametrima.[129] Prema tome, kako bi korpus bio reprezentativan, potrebno je definirati kriterije po kojima će se skupljati građa poput vrsta, razdoblja nastanka i dužine tekstova, osobina autora, žanra, medija (pisani ili govoreni jezik) itd.[130] S ciljem postizanja što bolje reprezentativnosti građe HNK je podijeljen u pet potkorpusa koji predstavljaju najzastupljenija područja pisanoga jezika: novine, časopisi, knjige, beletristika, eseji i govori.[131] Valja napomenuti da tekstovi koji su ulazili u sastav korpusa nisu digitalizirani ni na koji način, nego da su u obzir dolazili samo tekstovi nastali u digitalnom obliku.

Hrvatski elektronski tekstovni arhiv zbirka je tekstova koja je organizirana, no za koju ne vrijede ograničenja opsega, vremenskog raspona niti vrste teksta kao u slučaju korpusa. Svaka zbirka tekstova koju sadrži čini zaseban korpus, primjerice, sva djela Marka Marulića na hrvatskom jeziku.

---

[121] Tadić 2003., str. 87-89
[122] Klobučar Srbić 2008., str. 47
[123] Tadić 2003., str. 89
[124] Klobučar Srbić 2008., str. 47
[125] Tadić 2003., str. 91
[126] Tadić 1998., str. 3
[127] Tadić 2003., str. 89
[128] Hrvatski nacionalni korpus. (4.7.2015.).
[129] Tadić 1998., str. 1
[130] Tadić 1998., str. 5
[131] Klobučar Srbić 2008., str. 47

Glede zapisa korpusa, HNK-a obilježen je s pomoću XCES standarda te koristi standard UNICODE za kodiranje pismena. U upotrebi je i alat 2XML koji služi za pretvaranje drugih formata u XML.[132]

Od svojih početaka HNK je zamišljen kao jezični resurs koji će u cijelosti biti dostupan putem interneta.[133] Omogućen je ograničen pristup, jer se korpus konstantno nadopunjuje.[134] Što se tiče pretraživanja korpusa, trenutna verzija 3.0 omogućuje istodobnu pretragu pomoću više riječi, tj. upotrebu sintagmi, regularnih izraza, pretragu s pomoću vrste riječi, gramatičke kategorije, automatsko pronalaženje kolokacija te pretragu putem lema i gramatičkih kategorija što rezultira svim kombinacijama *prijedlog + opća imenica*.[135]

HNK bi mogao značajno napredovati i razvijati se u skladu sa svjetskim kretanjima, no problem koji se pritom najčešće javlja jest nedostatak sredstava.[136] Budući da je hrvatski jezik s relativno malim brojem govornika, ne može se očekivati da za jezične resurse na hrvatskom jeziku postoji jednak interes u gospodarskom smislu kao, primjerice, za resurse na engleskom ili njemačkom jeziku. Prema tome, država bi trebala snositi najveći dio troškova.

Ipak, računalna obrada jezika od velike je važnosti za Republiku Hrvatsku i njezine stanovnike – sadašnje i buduće pa stoga treba imati „status strateškog istraživanja za Republiku Hrvatsku“.[137]

3.3.2.3. Hrvatska jezična riznica

Nadalje, projekt *Hrvatska jezična riznica* pokrenut je 2005. godine na Institutu za hrvatski jezik i jezikoslovlje pod vodstvom Dunje Brozović Rončević.[138] Glavni cilj projekta jest izrada jezičnih resursa za hrvatski jezik koji će biti dostupni putem interneta. U okviru projekta prikuplja se građa iz svih područja znanosti koja u vremenskom smislu obuhvaća razdoblje od druge polovice 19. stoljeća pa sve do današnjih dana.[139] Prikupljena građa se po potrebi digitalizira, a upotrijebit će se za izradu reprezentativnog korpusa hrvatskoga standardnoga jezika koji je temelj za stvaranje Velikoga rječnika hrvatskoga jezika te za druga lingvistička istraživanja.[140]

---

[132] Tadić 2003., str. 91-93
[133] Tadić 2003., str. 87
[134] Hrvatski nacionalni korpus. (4.7.2015.).
[135] Korpus. // Hrvatski nacionalni korpus. (4.7.2015.).
[136] Tadić 2003., str. 91-100
[137] Tadić 1996., str. 607
[138] Klobučar Srbić 2008., str. 47
[139] Hrvatska jezična riznica. // Institut za hrvatski jezik i jezikoslovlje. (4.7.2015.).
[140] Klobučar Srbić 2008., str. 47

3.3.2.4. hrWac

Naposljetku, valja spomenuti najveći hrvatski korpus[141] *hrWac*. Riječ je o mrežnom korpusu koji u svojoj verziji 2.0 sadrži 2 milijarde pojavnica[142] koje su prikupljene s internetskih stranica domene .hr[143].

[141] Kukavica, V. Hrvatski jezik na Internetu : jezične tehnologije za hrvatski. // Zavod za kulturu vojvođanskih Hrvata. (4.7.2015.).
[142] Ljubešić, N. Upotreba jezičnih tehnologija u digitalizaciji teksta i njegovoj daljnjoj obradi. // Četvrti festival hrvatskih digitalizacijskih projekata. (4.7.2015.).
[143] Šojat; Srebačić; Štefanec 2013., str. 81

4. Računalni učenički korpusi

Računalni učenički korpusi mogu se definirati kao zbirka pisanih ili govornih tekstova u elektroničkom obliku pri čemu je bitno napomenuti da su autori tih tekstova neizvorni govornici dotičnog jezika.[144] Njihova je namjena višestruka – uglavnom se upotrebljavaju za praćenje učenja stranih jezika te za usporedbu jezične kompetencije neizvornih govornika s onom izvornih govornika.[145] Učenički korpusi temelje se na učenju jezika, tj. na usvajanju drugog[146] i stranog jezika. Istraživanje usvajanja stranih jezika ranije se većinom temeljilo na introspekciji nekoliko informanata[147] što kao metoda često daje ograničene rezultate, jer malen i nereprezentativan uzorak onemogućava donošenje općih zaključaka[148]. Prvi učenički korpusi uglavnom su sadržavali manje od 2000 pojavnica, a broj riječi često nije bio ni utvrđen, jer su se prije uporabe računala u ovom području riječi brojale ručno.[149] Veličina, raznolikost i računalna obrada[150] koja uključuje različite opcije korisne za istraživanja poput mogućnosti izračuna broja riječi u korpusu, sortiranja, obilježavanja, mogućnosti usporedbe, primjerice, osobitosti međujezika učenika kao jezične varijante s karakteristikama materinskog jezika[151] neke su od prednosti modernih učeničkih korpusa u odnosu na ranije korištene resurse prilikom istraživanja. Raniji korpusi uglavnom su se temeljili na analizi pogrešaka (tzv. *Error analysis*)[152]. Istraživači nisu uzimali u obzir različite okolnosti koje su mogle utjecati na učenikove rezultate. Korpusi su se svodili na prikaz učeničkih pogrešaka često i izvan konteksta, zanemarujući tako cjelovitu sliku o razini znanja nekog jezika. Stoga se javila potreba za novim metodama istraživanja kako bi se došlo do većeg broja kvalitetnijih podataka. Tako su današnji učenički korpusi u digitalnom obliku što otvara mogućnost uporabe različitih računalnih alata kako bi se ubrzao i olakšao postupak istraživanja. Tijekom vremena uvelike su uznapredovali u svojim mogućnostima glede lingvističke analize. Omogućuju napredno pretraživanje, praćenje različitih parametara unutar korpusa, poput izračuna broja riječi, usporedbe korpusa, uočavanje ponavljanja određenih jezičnih pogrešaka. Osim toga, današnji su učenički korpusi mnogo opsežniji što omogućuje različite jezične analize, a i njihova izrada podliježe puno većem broju strogih kriterija. Prilikom izrade

[144] Granger 2004., str. 124
[145] Granger 1998.
[146] Second language acquisition. // Wikipedia : the free encyclopedia. (4.7.2015.).
[147] Granger 1998.
[148] Granger 2004., str. 124
[149] Granger 1998.
[150] Granger 2004., str. 124
[151] Granger 2004., str. 127-128
[152] Error analysis. // Wikipedia : the free encyclopedia. (4.7.2015.).

korpusa veoma je važno pažljivo definirati kriterije te voditi računa o specifičnostima jezika, učenika i različitih načina odnosno okolnosti učenja jezika. Što se tiče karakteristika samog jezika, valja voditi računa o mediju odnosno je li riječ o pisanim ili govornim korpusima, o vrsti tekstova obuhvaćenih korpusom, o temi, stupnju stručnosti teksta te vrsti zadatka unutar kojeg je učenik stvorio određeni tekst. Glede pak autora tekstova unutar korpusa, tj. onih koji uče određeni strani jezik – osim spola i dobi, potrebno je uzeti u obzir faktore poput materinskog jezika, varijante materinskog jezika kojom učenik govori, tj. iz koje države odnosno regije potječe, govori li neki drugi strani jezik, koliko dugo uči jezik za koji se korpus izrađuje te je li ga učio kao strani jezik kod kuće ili za vrijeme boravka na govornom području dotičnog jezika. Svi navedeni faktori od velike su važnosti, jer u većoj ili manjoj mjeri definiraju kompetenciju odnosno uporabu dotičnog stranog jezika svakog učenika.[153]

4.1. Učenički korpusi u svijetu

Razvoj računalnih korpusa ove vrste započinje 90-ih godina prošlog stoljeća[154], kad velike nakladničke kuće te stručnjaci s područja učenja stranih jezika uočavaju prednosti računalnih učeničkih korpusa (CLC) u teoretskom i praktičnom smislu[155]. Tada dolazi do pokretanja raznih projekata u tom smjeru, a osobito je značajan projekt pod nazivom *International Corpus of Learner English* (ICLE) nastao na belgijskom Université Catholique de Louvain.[156] Riječ je o zbirci eseja koje su napisali studenti engleskog jezika i književnosti, dakle, korpus sadrži isključivo pisane tekstove. Treba spomenuti da se radi o studentima čiji se materinski jezici uglavnom razlikuju – od njemačkog, talijanskog i francuskog preko ruskog, bugarskog i češkog pa sve do kineskog, japanskog i hebrejskog.[157] Studentski eseji podijeljeni su u potkorpuse s obzirom na materinski jezik.[158] Projekt je nastao suradnjom između brojnih sveučilišta diljem svijeta što o profilu autora govori da engleski jezik studiraju u zemlji neengleskog govornog područja. Isto tako, valja napomenuti da je riječ o naprednom stupnju učenja engleskog jezika. Korpus je u stalnom porastu. Trenutno je u tijeku izrada treće verzije korpusa.[159]

Sastavljanje učeničkog korpusa mukotrpan je posao. Osim tekstova nastalih u digitalnom obliku, velik je broj tekstova koji još uvijek nastaju u tradicionalnom, rukopisnom obliku što

[153] Granger 1998.
[154] Mikelić Preradović; Berać; Boras 2015.
[155] Granger 1998.
[156] Mikelić Preradović; Berać; Boras 2015.
[157] ICLE. // Centre for English Corpus Linguistics. (4.7.2015.).
[158] Mikelić Preradović; Berać; Boras 2015.
[159] ICLE. // Centre for English Corpus Linguistics. (4.7.2015.).

uvelike otežava posao istraživačima, jer je takve tekstove potrebno najprije obraditi, tj. digitalizirati i pretvoriti u odgovarajući format. Obrada svih tekstova uključuje analizu pogreški. Pritom se misli na slučajne pogreške prilikom pisanja odnosno prepisivanja teksta, dok se učeničke jezične pogreške ne ispravljaju – one su od presudnog značenja za prikaz jezične kompetencije.[160]

Otkako je objavljen *International Corpus of Learner English*, dolazi do naglog razvoja učeničkih korpusa i za druge jezike. No, većim korpusima opsega od, primjerice, milijun pojavnica još uvijek se mogu pohvaliti samo veliki indoeuropski jezici poput njemačkog, španjolskog, talijanskog i francuskog.[161] Daleko najveći broj učeničkih korpusa izrađen je za engleski jezik. U skladu s tim, najpoznatiji korpusi danas u svijetu su *Longman Learner's Corpus, Cambridge Learner Corpus* te *Hong Kong University of Science and Technology Learner Corpus*[162]. Učeničke korpuse na engleskom jeziku slijede oni na njemačkom, španjolskom, francuskom i talijanskom. Što se tiče ostalih jezika, za finski jezik izrađena su 3, za arapski 2 korpusa. Prema popisu učeničkih korpusa u svijetu Centra za englesku korpusnu lingvistiku Katoličkog sveučilišta u Louvainu za kineski, nizozemski, estonski, irski, mađarski, korejski, norveški i švedski jezik postoji po jedan korpus te 9 višejezičnih korpusa s različitim jezičnim kombinacijama.[163]

Glede slavenskih jezika postoje samo tri učenička korpusa, i to za ruski (RULEC), slovenski (PiKUST) i češki jezik (CzeSL).[164]

RULEC odnosno *Russian Learner Corpus of Academic Writing* učenički je korpus ruskog jezika nastao kao rezultat suradnje između istraživača sa Sveučilišta u Portlandu te jednog od prestižnijih[165] ruskih sveučilišta National Research University – Higher School of Economics. Čine ga eseji koje je napisalo 36 američkih studenata ruskog na naprednom stupnju učenja jezika koji su ruski jezik učili kao ini jezik. Riječ je o relativno malom korpusu od oko 750 000 pojavnica, koji uključuje 3800 eseja.[166] Iz navedenog proizlazi da korpus nije pogodan za velika istraživanja općenitog karaktera, no s obzirom na veliki broj radova po pojedinom studentu može biti od višestrukog značaja, primjerice, za istraživanje razvoja učeničkog međujezika, etnografske studije itd.[167]

---

[160] Granger 1998.
[161] Mikelić Preradović; Berać; Boras 2015.
[162] Granger 1998.
[163] Learner corpora around the world. // Centre for English Corpus Linguistics. (4.7.2015.).
[164] Mikelić Preradović; Berać; Boras 2015.
[165] National Research University : Higher School of Economics. // Wikipedia : the free encyclopedia. (4.7.2015.).
[166] Mikelić Preradović; Berać; Boras 2015.
[167] RULEC : Russian Learner Corpus of Academic Writing. // Web corpora. (4.7.2015.).

PiKUST je sirovi korpus pisanih tekstova na slovenskom jeziku koji uključuje 35 000 pojavnica odnosno 128 eseja.[168] Pisali su ih učenici različitog porijekla, većinom na visokom stupnju jezične kompetencije (stupanj C1)[169] pa tako među njima nalazimo 18 različitih materinskih jezika. Bitno je napomenuti da je velika većina autora radova u korpusu, tj. njih oko 90% potječe s područja bivše Jugoslavije[170] što je takoreći očekivana brojka, jer 89% stranaca u Sloveniji potječe iz neke od država bivše Jugoslavije. Zatim je brojčano najviše izvornih govornika ruskog, engleskog, slovačkog itd. No, broj radova autora s ostalim jezicima kao materinskim većinom je zanemariv i teško se može uzeti kao relevantan materijal za istraživanje[171], jer može dovesti do pogrešnih rezultata[172]. Prema tome, može se reći da se radi o neuravnoteženom korpusu na temelju kojeg nije moguće donositi opće zaključke, no koji će ipak biti od višestruke koristi za jezična istraživanja.[173]

CzeSL je korpus češkog kao stranog jezika nastao u okviru projekta strukturnih fondova EU i češke vlade. Sastoji se od tri potkorpusa: Prvi sadrži eseje neizvornih govornika češkog na različitim stupnjevima učenja češkog jezika, drugi eseje stranih studenata čeških sveučilišta na diplomskom ili postdiplomskom studiju i treći sadrži školske eseje učenika romskog porijekla.[174]

4.1.1. Pisani korpusi

Jedna od presudnih odluka prilikom izrade korpusa jest odluka o tome hoće li sadržavati tekstove u pisanom obliku, transkripcije govornog jezika ili oboje. Iako korpusi govornog jezika iziskuju daleko više rada, gotovo je isključivo na temelju takvih korpusa, prema Johnu Sinclairu, zapravo moguće uvidjeti stvarno stanje jezika, jer korpusi govornog jezika sami po sebi sadrže jezik koji je u rijetkim slučajevima prethodno smišljen kao što je slučaj s tekstovima unutar korpusa pisanog jezika.[175]

Projekt pod nazivom *Kolipsi* možemo smatrati primjerom za ovu vrstu korpusa. Riječ je o projektu u okviru kojeg se istražuju lingvistički, sociolingvistički i socijalno-psihološki

[168] Mikelić Preradović; Berać; Boras 2015.
[169] Stritar 2009., str. 138
[170] Mikelić Preradović; Berać; Boras 2015.
[171] Stritar 2009., str. 137
[172] Granger 2004., str. 124
[173] Stritar 2009., str. 144
[174] The CzeSL-plain corpus. // Institute of the Czech National Corpus. (4.7.2015.).
[175] Sinclair 1991., str. 16

aspekti učenja drugog jezika[176] kod učenika u dobi od 17-18 godina.[177] Bitno je napomenuti da se istraživanje odnosi na područje Južnog Tirola koje je vrlo specifično glede jezičnih varijanti koje su u uporabi. Južni Tirol najsjevernija je talijanska pokrajina koja graniči s austrijskim saveznim zemljama Tirolom i Salzburgom te švicarskim kantonom Graubündenom (u kojem su tri službena jezika). Iz takvog geografskog položaja proizašla je jezična raznolikost, bogatstvo različitih jezika i nebrojenih dijalekata. Njemački je materinski jezik za 62% stanovništva Južnog Tirola, talijanski za njih 23% i ladinski (iz retoromanske skupine jezika) za 4% stanovništva.[178] Glede poznavanja drugog jezika, rezultati istraživanja pokazali su se poražavajućima – prolaznost na ispitu znanja drugog jezika bila je manja od 40% iako je riječ o jezicima koje ispitanici svakodnevno upotrebljavaju te žive na istom području. Projekt istražuje povezanost jezične kompetencije i vanjskih faktora koji imaju veliki utjecaj na učenje i uporabu drugog jezika s ciljem poboljšavanje jezične politike Južnog Tirola te poticanja učenja drugog jezika i višejezičnosti općenito.[179]

Nadalje, još jedan primjer pisanih korpusa jest *MeLLANGE Learner Translator Corpus* (*LTC*) nastao u okviru projekta pod nazivom *MeLLANGE (Multilingual eLearning in Language Engineering).* Projekt je nastao na temelju suradnje između sveučilišta i tvrtki iz Francuske, Austrije, Češke, Njemačke, Italije, Španjolske, Švicarske i Ujedinjenog Kraljevstva.[180]

Riječ je o višejezičnom korpusu sastavljenom od prijevoda tekstova iz različitih stručnih područja poput prava, tehnike, administrativnih i novinskih tekstova. Bitno je napomenuti da su prijevode izrađivali profesionalni prevoditelji, ali i studenti. Korpus je obilježen, tj. sadrži metapodatke i lingvističke informacije, ali su istaknute i prevoditeljske pogreške – smisaone i jezične. Ciljevi ovog projekta jesu uočavanje bitnih karakteristika tekstova iz navedenih područja i različitih fenomena koji se javljaju prilikom prevođenja, stvaranje jezičnih resursa za učenje jezika uključenih u projekt te stvaranje resursa koji će biti od pomoći prevoditeljima, ali i studentima i njihovim profesorima.[181]

[176] Valja razlikovati pojam 'drugi jezik' (second langugage, L2) od pojma 'strani jezik'. Drugi jezik je jezik koji osoba uči odnosno govori pored materinskog jezika, s tim da joj je nužno potreban za svakodnevnu komunikaciju, jer se, primjerice, radi o jeziku zemlje u kojoj osoba stanuje ili jedan od roditelja govori samo dotični jezik. O ovoj temi još će biti govora u nastavku.

[177] Kolipsi : die Südtiroler Schüllerinnen und die Zweitsprache : eine linguistische und sozialpsychologische Untersuchung. // EURAC. (4.7.2015.).

[178] Südtirol. // Wikipedia : die freie Enzyklopädie. (4.7.2015.).

[179] Kolipsi : die Südtiroler Schüllerinnen und die Zweitsprache : eine linguistische und sozialpsychologische Untersuchung. // EURAC. (4.7.2015.).

[180] About MeLLANGE. // MeLLANGE. (4.7.2015.).

[181] The MeLLANGE Learner Translator Corpus. // MeLLANGE. (4.7.2015.).

4.1.2. Govorni korpusi

Govorni korpusi sastoje se od zvučnih i tekstualnih zapisa. Postoje dvije vrste govornih korpusa: zvučni zapisi čitanog teksta i spontani govor. Zvučni zapisi čitanog teksta obuhvaćaju dijelove knjiga, televizijske vijesti, popise riječi itd. Spontani govor uključuje dijaloge, pripovijesti (osoba pripovijeda neku priču), zadatke u kojima, primjerice, jedna osoba drugoj objašnjava određenu rutu na karti ili zadatke u kojima dvije osobe dogovaraju vrijeme međusobnog susreta.[182] Prikupljanje građe za govorni korpus iznimno je zahtjevno. U nedostatku softvera za automatsko prepoznavanje govora, zvučne je zapise nužno ručno transkribirati što iziskuje mnogo vremena.[183]

Zanimljiv primjer ove vrste korpusa jest *The Giessen – Long Beach Chaplin Corpus* (*GLBCC*) izrađen za engleski jezik. Sastoji se od transkribiranih razgovora između izvornih govornika i onih koji uče engleski kao drugi ili kao strani jezik. Projekt je osmišljen na sljedeći način: Studenti u Kaliforniji – izvorni govornici i govornici engleskog kao drugog jezika te oni u njemačkom Giessenu koji uče engleski kao strani jezik podijeljeni su u parove i trebali su pogledati prvi dio nijemog filma *Useljenik* s Charlieom Chaplinom[184]. Nijemi je film odabran s namjerom kako bi sam film što manje utjecao na jezik sudionika.[185] Nakon toga jedan bi sudionik ispričao radnju filma do tog trenutka, dok bi drugi pogledao ostatak filma i prepričao svom partneru. Na kraju bi sudionici u parovima raspravljali o nekim aspektima filma. Na projektu je ukupno je sudjelovao 191 govornik i snimljeno 22 sata i 20 minuta zvučnog materijala.[186] Glede strukture sudionika, većina izvornih govornika su Amerikanci, a većina neizvornih Nijemci, no među neizvornim govornicima nailazimo i sudionike s drugim korijenima. Cilj projekta bio je slobodno izražavanje kroz prepričavanje odnosno interpretaciju filma te opisati razlike između govora, tj. jezika izvornih i neizvornih govornika engleskog jezika.[187] U tom smislu, posebna je pozornost posvećena diskursnim oznakama[188], tj. riječima ili uzrečicama kojima govornik povezuje rečenice, jer one najčešće razlikuju govor izvornih i neizvornih govornika nekog jezika.

[182] Speech corpus. // Wikipedia : the free encyclopedia. (4.7.2015.).
[183] Granger 2004., str. 125
[184] Mukherjee, J. Corpus linguistics and language pedagogy: The state of the art – and beyond. // Universität Giessen. (4.7.2015.).
[185] Müller 2005., str. 31-34
[186] GLBCC. // University of Oxford Text Archive. (4.7.2015.).
[187] Müller 2005., str. 31-34
[188] Pranjković, I. Jezikoslovlje i tekst. // Vijenac. (4.7.2015.).

4.2. Računalni učenički korpus hrvatskog jezika

4.2.1. Materinski jezik, drugi i strani jezik

U novije vrijeme, osim pojma „strani jezik“ sve češće je u uporabi pojam „drugi jezik“. Ovdje valja najprije razgraničiti ova dva pojma. Drugi jezik je jezik koji osoba uči odnosno govori osim materinskog jezika, s tim da joj je nužno potreban za svakodnevnu komunikaciju, jer se, primjerice, radi o jeziku zemlje u kojoj osoba stanuje ili jedan od roditelja govori samo dotični jezik iz čega proizlazi ključan aspekt usvajanja drugog jezika – usvajanje jezika unutar nekog oblika zajednice koja govori dotičnim jezikom. A stranim se jezikom smatra svaki jezik koji se usvaja nakon materinskog, i to na području na kojem se ne govori dotični jezik.[189] Do razgraničavanja ovih pojmova dolazi, jer bitna razlika između njih zbog današnjeg načina života, globalizacije, češćih putovanja, ali i migracija s ciljem trajnog nastanjivanja sve više dolazi do izražaja. Pod nazivom „prvi jezik“ podrazumijeva se prvi jezik koji osoba u životu usvoji, tj. materinski jezik. Materinski se jezik usvaja na specifičan način i osoba ga upotrebljava ne razmišljajući, intuitivno. Drugi jezik odražava neke sličnosti s materinskim jezikom glede usvajanja samog jezika, jer se i drugi jezik uglavnom usvaja na takoreći prirodan način – kroz druženje s djecom, u vrtiću, za razliku od stranog jezika koji se uglavnom usvaja u školi, na različitim jezičnim tečajevima te učenici često nisu u prilici naučeni jezik upotrebljavati u nekoj svakodnevnoj, prirodnoj situaciji koja nije unaprijed „osmišljena“. Razlikovanje ova dva pojma osobito je značajno u sociolingvističkom smislu, u kontekstu učenja stranih jezika pri čemu su utjecaji okoline i drugih jezika koje su učenici usvajali u nekom razdoblju svog života od iznimne važnosti za učenje novog jezika.[190] U stručnoj se literaturi sve češće susreće i izraz *ini jezik* koji objedinjuje gore navedene mogućnosti te, prema tome, značenje ovog izraza glasi drugi ili strani jezik.[191]

A kako bi jezična kompetencija onih koji uče neki strani jezik bila što transparentnije i preciznije opisana i kako bi se omogućila jednostavnija komunikacija i suradnja među stručnjacima na tom području, Vijeće Europe stvorilo je *Zajednički europski referentni okvir za jezike* koji daje detaljne smjernice za poučavanje modernih jezika te definira stupnjeve jezične kompetencije u različitim fazama učenja jezika. Neki od ciljeva Zajedničkog europskog okvira su održavanje kvalitete učenja stranih jezika na visokom nivou, poticanje

[189] Second language. // Wikipedia : the free encyclopedia. (4.7.2015.).
[190] Jelaska 2005., str. 24-37
[191] Macan; Kolaković 2008. str. 34

cjeloživotnog učenja stranih jezika, poštovanje kulturne raznolikosti itd.[192] Opis stupnjeva ZEROJ-a razumljiv je i koristan onima koji uče ili žele učiti pojedini jezik, no služi i kao dobra orijentacija učiteljima za organizaciju nastave i praćenje napretka učenika.
Stupnjeve učenja jezika možemo podijeliti na tri temeljna stupnja: A – temeljni korisnik, B – samostalni korisnik i C – iskusni korisnik. Najčešća je podjela na šest stupnjeva koja proizlazi iz razrade ovih triju stupnjeva:

Temeljni korisnik	A1	Učenik razumije i upotrebljava svakodnevne izraze i jednostavne rečenice. Znade predstaviti sebe i druge te odgovarati na pitanja te vrste. Može voditi jednostavan razgovor ukoliko sugovornik govori polako i razgovijetno te je spreman pomoći.
	A2	Učenik razumije izolirane rečenice i česte izraze vezane uz područja od neposrednog osobnog interesa. Može se sporazumjeti u jednostavnim, svakodnevnim situacijama. Može jednostavnim riječima govoriti o svojem obrazovanju, neposrednoj okolini te o stvarima vezanim uz neposredne potrebe.
Samostalni korisnik	B1	Učenik razumije glavne misli, ukoliko se upotrebljava standardni jezik i ukoliko se govori o poznatim temama vezanim uz posao, školu, neposredne potrebe. Može jednostavnim riječima govoriti o poznatim temama, o svojim doživljajima i područjima koja ga zanimaju.
	B2	Učenik razumije glavne točke složenih tekstova vezanih uz konkretne i apstraktne teme te stručne rasprave iz područja svoje struke. Može govoriti tečno i spontano i bez poteškoća razgovarati s izvornim govornicima. U stanju je jasno i detaljno se izražavati o velikom broju različitih tema.
Iskusni korisnik	C1	Učenik razumije duže, zahtjevne tekstove iz velikog broja različitih područja te može prepoznati implicitna značenja. Jasno, spontano i opširno se izražava o složenim temama. Može učinkovito i bez poteškoća upotrebljavati jezik u društvenim i poslovnim situacijama.
	C2	Učenik bez napora razumije sve što čuje ili pročita i u stanju je jasno sažeti i prenijeti informacije iz različitih izvora. Izražava se tečno i precizno, shvaća i razlikuje i sitne nijanse u značenju.[193]

Tablica br. 1 Stupnjevi ZEROJ-a

[192] Čeliković, V. (ur.) 2005., str 1-8
[193] Čeliković, V. (ur.) 2005., str. 24

Nastava na *Croaticumu*, Centru za hrvatski kao drugi strani jezik Filozofskog fakulteta u Zagrebu, također je organizirana u okviru šest stupnjeva koji su glede razine jezične kompetencije ekvivalentni stupnjevima ZEROJ-a:

1. 1A (usporediv s A2+)
2. 1B (usporediv s B1)
3. 2A (usporediv s B1+)
4. 2B (usporediv s B2)
5. 3A (usporediv s B2+)
6. 3B (usporediv s C1)[194]

4.2.2. O korpusu općenito

Računalni učenički korpus hrvatskog jezika prvi je računalni korpus ove vrste za hrvatski jezik. Korpus sačinjavaju autentični učenički eseji prikupljeni u okviru tečajeva hrvatskog jezika na *Croaticumu*.[195]

Korpus u trenutku pisanja ovog rada sadrži milijun pojavnica, a računalnom obradom korpusa koja će uslijediti analizirat će se različiti aspekti učeničkog jezika te odstupanja od standarda. Na korpusu će se primijeniti različiti načini obilježavanja poput označavanja vrsta riječi (tzv. *part-of-speech tagging*), vrsta učeničkih pogrešaka (*error tagging*) i značenjskog označavanja te tako omogućiti istraživanje međujezika[196] učenika i načina kako im olakšati usvajanje hrvatskog jezika što će u konačnici omogućiti izradu različitih računalnih pomagala za interaktivno učenje hrvatskog na različitim stupnjevima jezične kompetencije, kao i njihovu javnu dostupnost. Potencijalni korisnici korpusa bili bi učenici i profesori hrvatskog kao inog jezika u Hrvatskoj i inozemstvu, hrvatski iseljenici, useljenici u Hrvatskoj, znanstvenici i stručnjaci iz različitih područja i drugi.[197]

4.2.3. Hrvatski inojezični korpus tekstualnih zapisa (HINKOT)

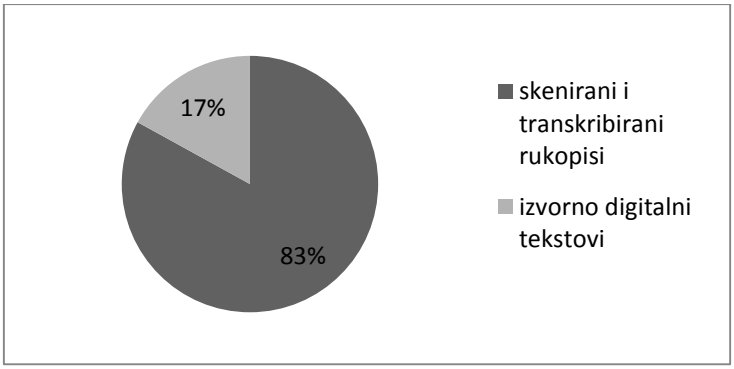
HINKOT korpus sadrži pisane tekstove učenika koji uče hrvatski kao strani jezik te je obilježen iscrpnim sociolingvističkim metapodacima: spol, dob, materinski jezik, poznavanje

[194] Skraćeni semestralni tečaj. // Croaticum. (4.7.2015.).
[195] Mikelić Preradović; Berać; Boras 2015.
[196] Prilikom usvajanja drugog ili stranog jezika formira se tzv. *međujezik*. Riječ je o jezičnoj varijanti koja funkcionira kao sustav u kojem vrijede određena pravila. No, svojstvena su mu i odstupanja koja su dokaz jezičnog razvoja (odstupanje ≠ pogreška). S napretkom učenika u učenju inog jezika i međujezik se mijenja, broj odstupanja se smanjuje. (Macan; Kolaković 2008., str. 34-35)
[197] Mikelić Preradović; Berać; Boras 2015.

drugih stranih jezika, duljina boravka u Hrvatskoj, trenutna razina jezične kompetencije na hrvatskom jeziku, vrijeme i način usvajanja hrvatskoga jezika, ostali kontakti s hrvatskim jezikom, materinski jezik roditelja. Tijekom akademskih godina 2013./2014. te 2014./2015. prikupljeni su eseji učenika sa svih šest stupnjeva učenja jezika. Rukopisni tekstovi digitalizirani su prije i nakon ispravljanja pogrešaka kako bi se mogli kasnije upotrijebiti u svrhu obilježavanje pogrešaka u korpusu te za detaljnu raščlambu i analizu učeničkih pogrešaka. Digitalni su eseji također prikupljani prije i poslije korekture. Bitno je istaknuti da pogreške imaju različitu težinu s obzirom na stupanj učenja jezika, tj. kao pogreške se računaju samo odstupanja koja ulaze u nastavne sadržaje koji su do trenutka pisanja eseja trebali biti usvojeni. I dužina eseja varira s obzirom na stupanj učenja jezika.

U danom trenutku još sirovi korpus sastoji se od transkribiranih rukopisnih eseja u izvornom obliku, tj. eseji sadrže i ispravke samih učenika poput, primjerice, umetanja riječi ili rečenica, promjene redoslijeda riječi, precrtanih riječi ili rečenica itd. Nakon što su svi eseji prikupljeni u digitalnom obliku, pretvoreni su u format XML i spremljeni. Profesori *Croaticuma* dostavljali su i podatke o esejima poput teme, jezične zahtjevnosti, uvjeta pod kojim su nastali (u okviru domaćeg zadatka, ispita, koliko je vremena stavljeno na raspolaganje za određeni zadatak i sl.) itd. Također, prikupljani su i upitnici koje su polaznici ispunjavali, a iz kojih su dobiveni sociološki podaci poput spola, dobi, stupnja učenja jezika, materinskog jezika itd.

HINKOT se u travnju 2014. sastojao od ukupno 2834 skenirana teksta (1417 izvornih tekstova i njihovih ispravljenih verzija), 1417 transkripata u RTF formatu iz kojih su uklonjeni osobni podaci polaznika i koje će se pretvoriti u format XML te 290 eseja izvorno nastalih u digitalnom formatu. Iz priloženog grafikona jasno je vidljivo da među učeničkim esejima dominiraju rukopisi što je znatno povećalo opseg posla istraživačima.



Graf br.1 Omjer rukopisnih i izvorno digitalnih tekstova u HINKOT-u

HINKOT je u travnju 2014. sadržavao ukupno 301 697 riječi u esejima od ukupno 295 polaznika s 36 različitih materinskih jezika.

U lipnju 2015. korpus je sadržavao 5668 skeniranih tekstova (učenički original i lektorska korekcija), 2834 prepisana eseja te 866 eseja u izvorno digitalnom obliku te je tako dobiven korpus veličine 602 096 riječi. Preostalih 1177 skeniranih eseja potrebno je prepisati, što će u konačnici rezultirati korpusom od milijun pojavnica.

4.2.4. Hrvatski inojezični korpus akustičkih zapisa (HINKAZ)

HINKAZ je još uvijek sirov korpus koji čine snimke čitanih tekstova i zvučnih zapisa spontanoga govora. Teorijske postavke izrade ovog korpusa postavila je Banković-Mandić (2012) u doktorskoj disertaciji *Izgovorna obilježja učenika hrvatskoga kao drugoga i stranoga jezika na različitim stupnjevima znanja*. Ovaj rad tematizira razvoj izgovora hrvatskog kao inog jezika na različitim stupnjevima učenja. Nastao je na temelju analize zvučnog korpusa sastavljenog od 121 zvučnog zapisa neizvornih govornika hrvatskog jezika iz 24 zemlje i dva zapisa izvornih govornika. Svaki je govornik trebao pročitati isti tekst i opisati istu sliku, a njihov izgovor odnosno fonološka i fonetska odstupanja ocjenjivali su stručnjaci, ali i nestručni procjenitelji. Također, izgovor hrvatskog jezika analiziran je s obzirom na materinske jezike govornika te različite stupnjeve učenja jezika.[198]

Što se tiče HINKAZ-a, u zimskom semestru akademske godine 2013./2014. prikupljeno je 9 sati i 25 minuta zvučnih zapisa od 144 učenika. Cilj korpusa jest omogućiti sustavnu analizu pogrešaka prilikom izgovora hrvatskog jezika kod govornika različitog porijekla, tj. s različitim materinskim jezikom. Posebno će se uzeti u obzir izvorni govornici te učenici čiji je materinski jezik neki od jezika srodnih hrvatskome. U sljedećoj fazi istraživanja transkribirat će se zvučni zapisi, povezat će ih se s HINKOT-om te ih pripremiti za proces obilježavanja pogrešaka. Naposljetku bi HINKAZ trebao pružati mogućnost pretrage po različitim parametrima poput materinskog jezika, stupnja učenja hrvatskoga itd. [199]

[198] Banković-Mandić, I. Izgovorna obilježja učenika hrvatskoga kao drugoga i stranoga jezika na različitim stupnjevima znanja. // Hrvatska znanstvena bibliografija. (18.8.2015.).
[199] Mikelić Preradović; Berać; Boras 2015.

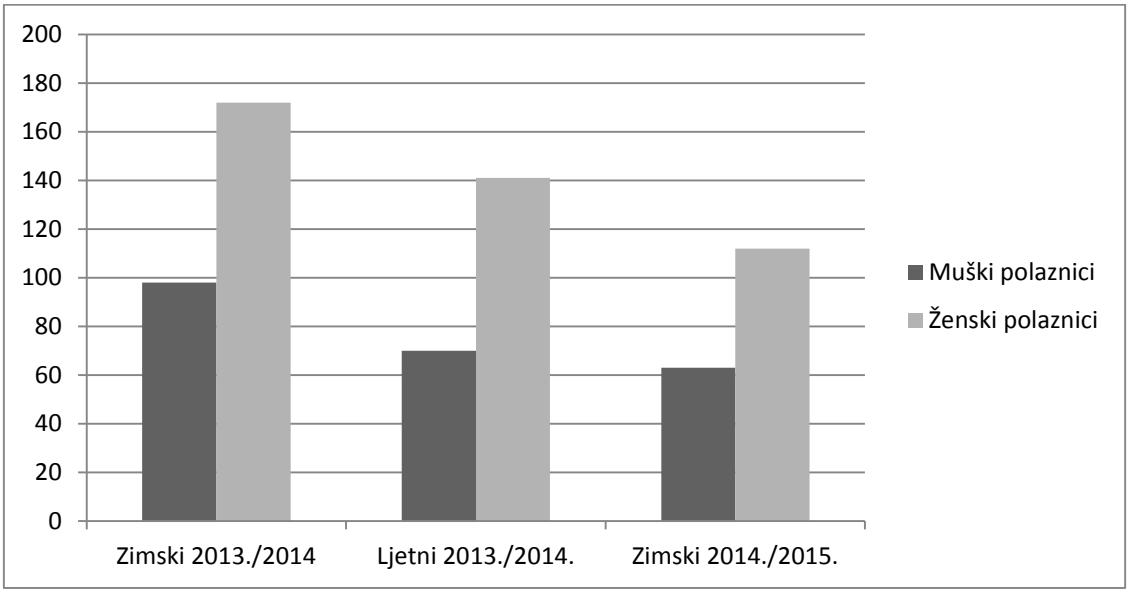
5. Istraživanje

Istraživanje prati polaznike *Croaticuma* kroz tri semestra: od zimskog 2013./2014. preko ljetnog 2013./2014. do zimskog semestra 2014./2015. Prvi dio istraživanja obuhvaća sociolingvističke podatke o polaznicima, dok je u drugom dijelu dan pregled tema koje učenici najčešće obrađuju u svojim esejima te analiza tema i njihove složenosti s obzirom na stupanj učenja jezika. Sociolingvistički podaci ekstrahirani su iz upitnika koje su učenici ispunjavali uglavnom na početku semestra, a podaci o temama eseja dobiveni su iz samih eseja – digitaliziranih ili izvorno nastalih u digitalnom obliku te iz upitnika o temi koji su popunjavali nastavnici. Rezultati su popraćeni grafičkim prikazom.

5.1. Sociolingvistički podaci o polaznicima

5.1.1. Spol

Spolna struktura polaznika konstantna je kroz promatrano vremensko razdoblje. Kao polaznici dominiraju ženske osobe koje kroz tri semestra čine više od 60% polaznika, a u ljetnom semestru 2013./2014. čak 67%. Što je uzrok tomu? Jesu li žene sklonije učenju jezika? Može se reći da su žene sklonije studiju jezika, muškarci čine manjinu na studijima humanističkih znanosti općenito. Nameće se zanimljivo sociološko pitanje o tome mogu li se navedene činjenice smatrati društveno odnosno okolinski uvjetovanim ponašanjem ili pak ponašanjem uvjetovanim prirodnim sklonostima, sposobnostima.[200]

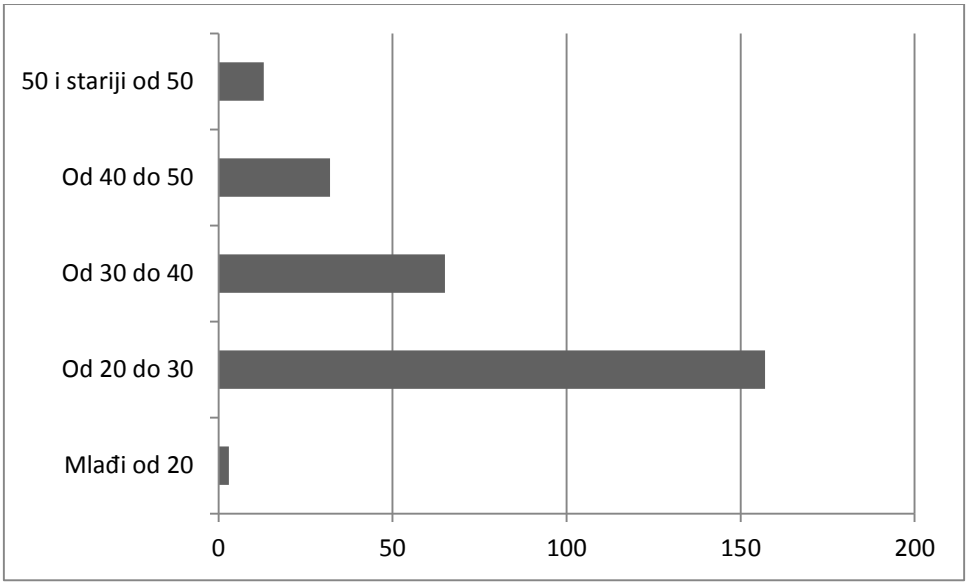


Graf br. 2 Spolna struktura polaznika

[200] Više djevojaka upisuje studij i više ih završava. // Hrvatska radiotelevizija. (5.9.2015.)

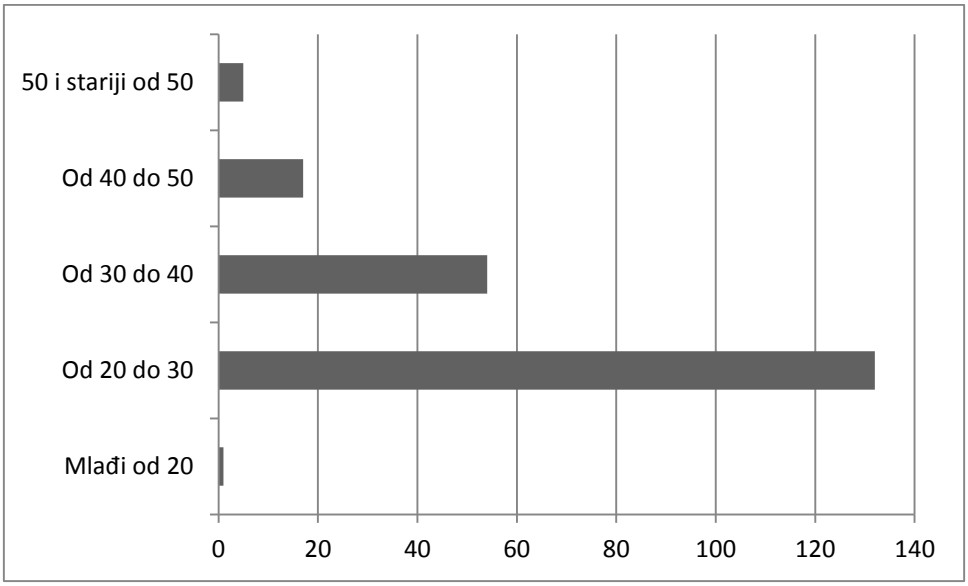
5.1.2. Dob

Dobna struktura polaznika analizirana je za svaki semestar posebno. Tako u zimskom semestru 2013./2014. veliku većinu, čak 82%, čine polaznici u dobi od 20 do 40 godina, a među njima najveći je broj polaznika starijih od 20 a mlađih od 30 godina – čak 58%. Najmanji je broj polaznika mlađih od 20 godina. S obzirom na jasnu dominaciju navedenih dobnih skupina, zanimljivo je istaknuti da je 5% polaznika navršilo pedesetu godinu života



Graf br.3 Zimski semestar ak. god. 2013./2014.

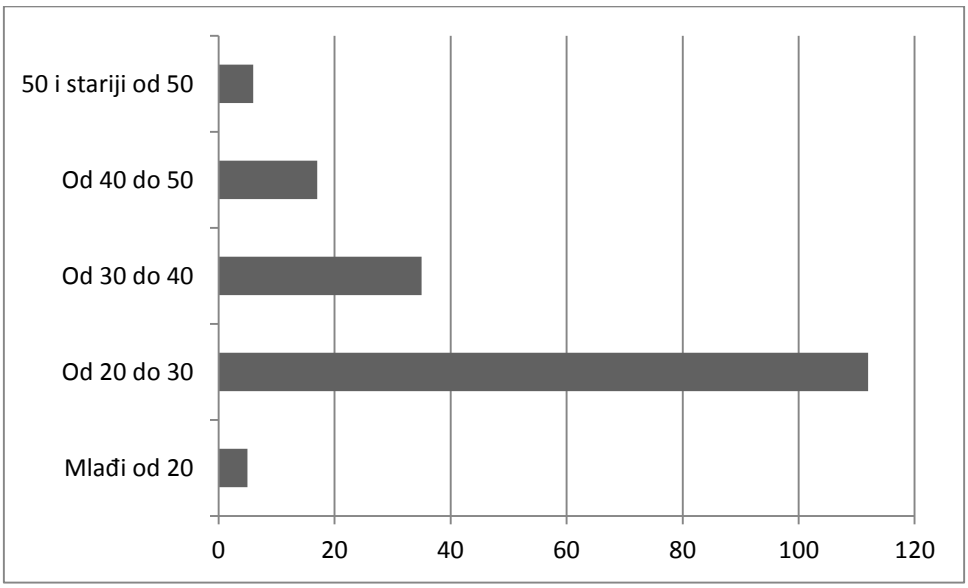
Što se tiče ljetnog semestra 2013./2014., dominanta je dobna skupina polaznika od 20 do 30 godina s udjelom od 63% polaznika. Slijedi je skupina polaznika u dobi od 30 do 40 godina s 26%. Iako u znatno manjem broju od prethodne dvije skupine, valja istaknuti da su i u ovom semestru tečaj hrvatskog jezika pohađali i polaznici mlađi od 20 te stariji od 40 godina.



Graf br.4 Ljetni semestar ak. god. 2013./2014.

I na koncu, u zimskom semestru akademske godine 2014./2015. – kao i u prethodna dva – najmanja je skupina polaznika mlađih od 20 godina i starijih od 40 godina, a polaznici u dobi od 20 do 30 godina čine veliku većinu odnosno 64% polaznika. Značajan udio čine i polaznici između 30 i 40 godina – 20%.

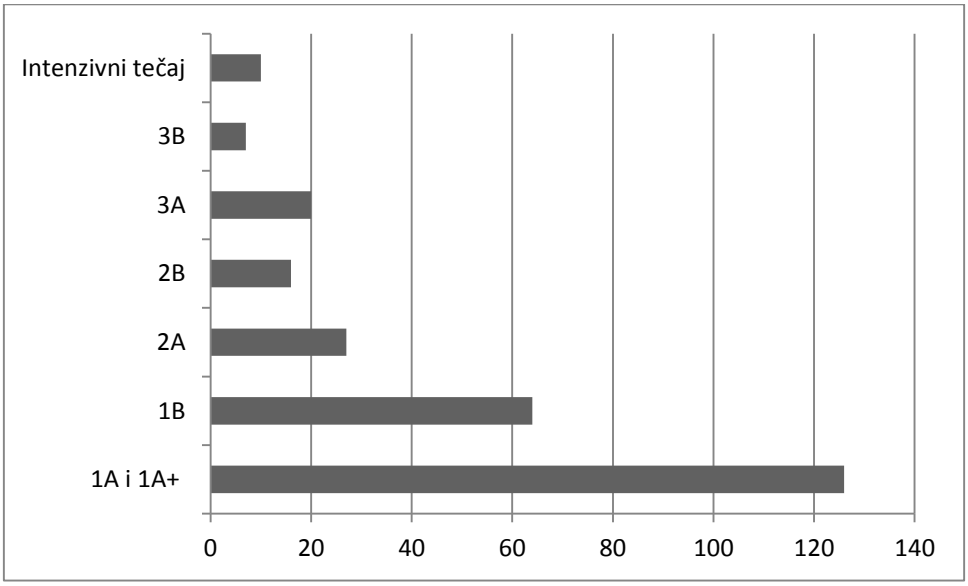
Uzimajući u obzir sva tri semestra, najstariji polaznik ima 68, najmlađi 16 godina.



Graf br.5 Zimski semestar ak. god. 2014./2015.

5.1.3. Stupanj

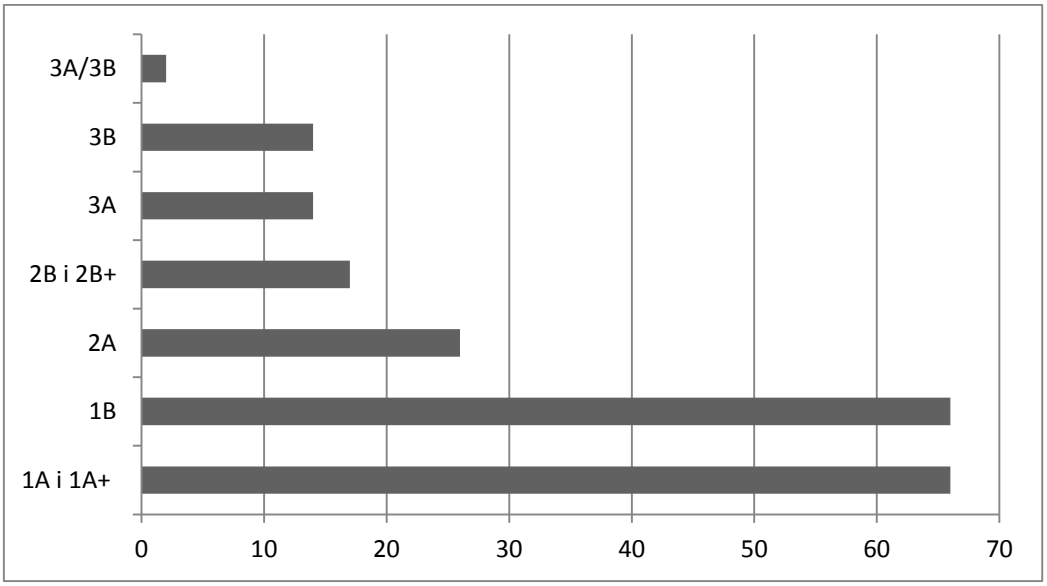
Pod pretpostavkom da podaci o polaznicima kojima raspolažemo odgovaraju stvarnom stanju, može se reći da najviše polaznika hrvatski jezik uči na početnom stupnju te kako stupnjevi idu prema naprednijem, smanjuje se broj polaznika. Tako u zimskom semestru 2013./2014. polaznici tečaja za prva dva stupnja čine 71% od ukupnog broja svih polaznika, a polaznici naprednog tečaja svega 9%. Srednji stupnjevi 2A i 2B zastupljeni su s udjelom od 16%. U ovaj prikaz uključeni su i polaznici intenzivnog tečaja u okviru *Croaticuma* koji je usmjeren na prevoditelje i sve one koji se profesionalno bave hrvatskim jezikom, ali i na sve ostale čije je poznavanje hrvatskoga jezika na razini B1 ili višoj. Ovaj tečaj, za razliku od ostalih u sklopu *Croaticuma* koji su uglavnom koncipirani semestralno, traje samo dva tjedna te obrađuje teme poput ekonomije, prava, politike[201] itd.



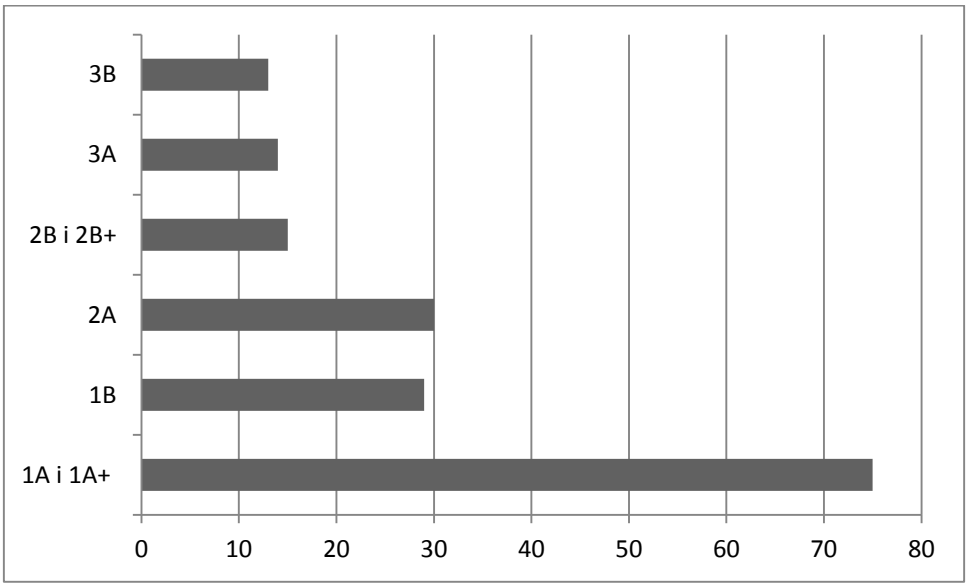
Graf br.6 Zimski semestar ak. god. 2013./2014.

Isti trend nastavlja se i u ljetnom semestru 2013./2014. i u zimskom 2014./2015. pa tako 64% odnosno 59% od ukupnog broja polaznika čine polaznici na početnim stupnjevima učenja jezika. Srednji stupnjevi čine 21% odnosno 26%, dok su napredni stupnjevi 3A i 3B i u ljetnom 2013./2014. i u zimskom semestru 2014./2015. zastupljeni udjelom od 15%.

[201] Specijalizirani intenzivni tečaj. // Croaticum. (4.7.2015.).



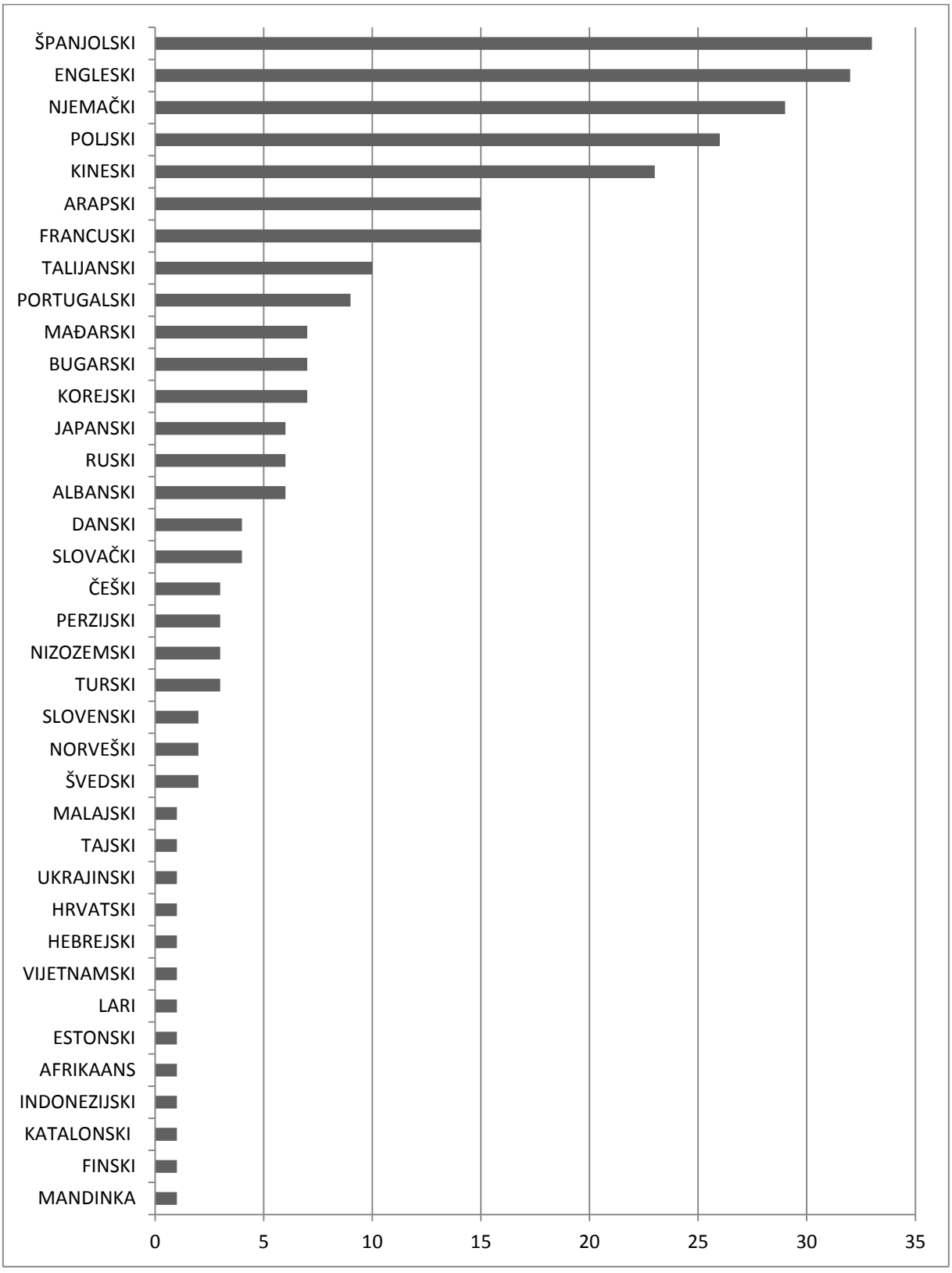
Graf br.7 Ljetni semestar ak. god. 2013./2014.



Graf br.8 Zimski semestar ak. god. 2014./2015.

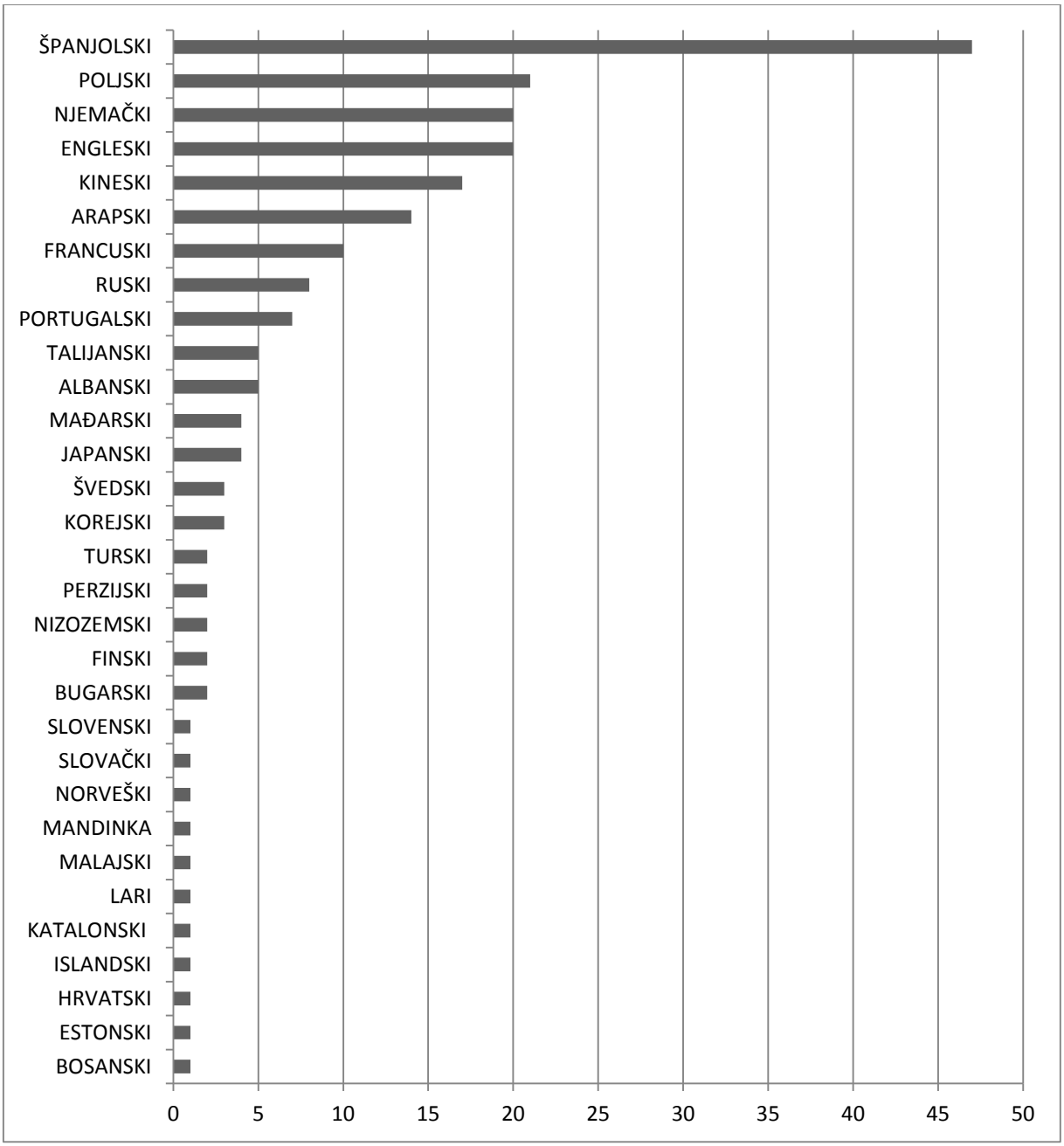
5.1.4. Materinski jezik

Dan je pregled materinskih jezika polaznika za svaki semestar zasebno. U zimskom semestru susrećemo 37 materinskih jezika s različitim postotkom zastupljenosti među polaznicima.



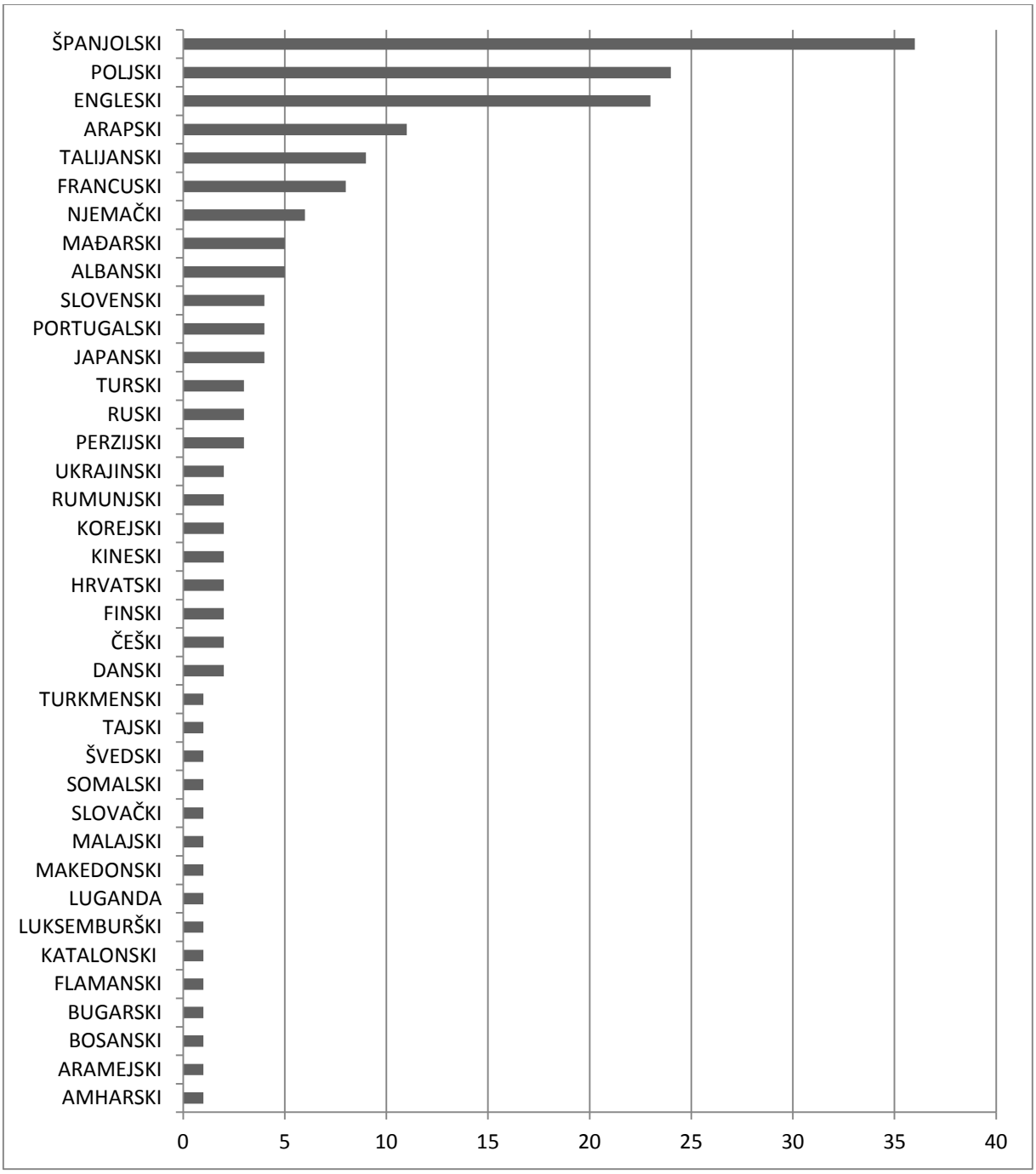
Graf br. 9 Zimski semestar ak. god. 2013./2014.

Među polaznicima u ljetnom semestru 2013./2014. razlikujemo 31 materinski jezik.



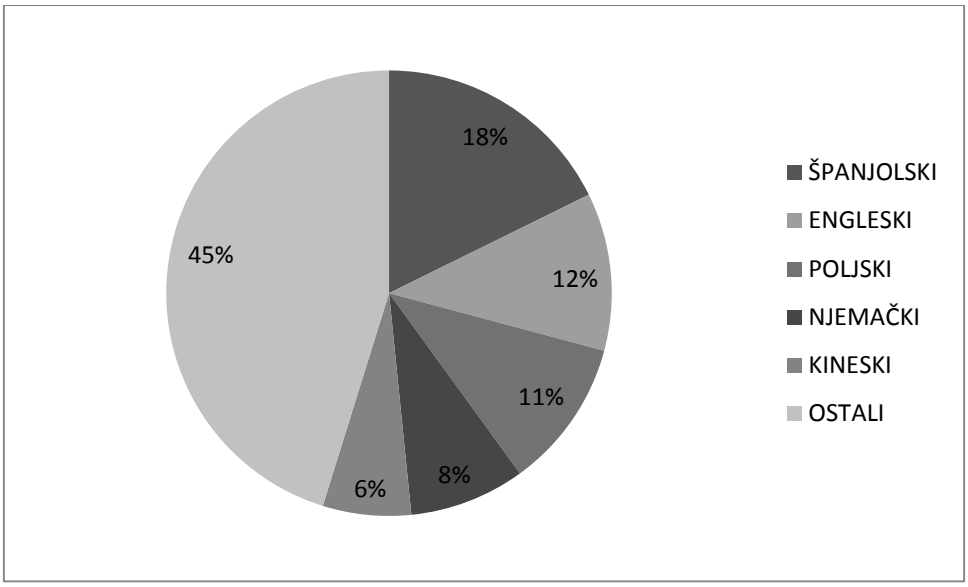
Graf br. 10 Ljetni semestar ak. god. 2013./2014.

U zimskom semestru 2014./2015. nailazimo na 38 različitih materinskih jezika.



Graf br. 11 Zimski semestar ak. god. 2014./2015.

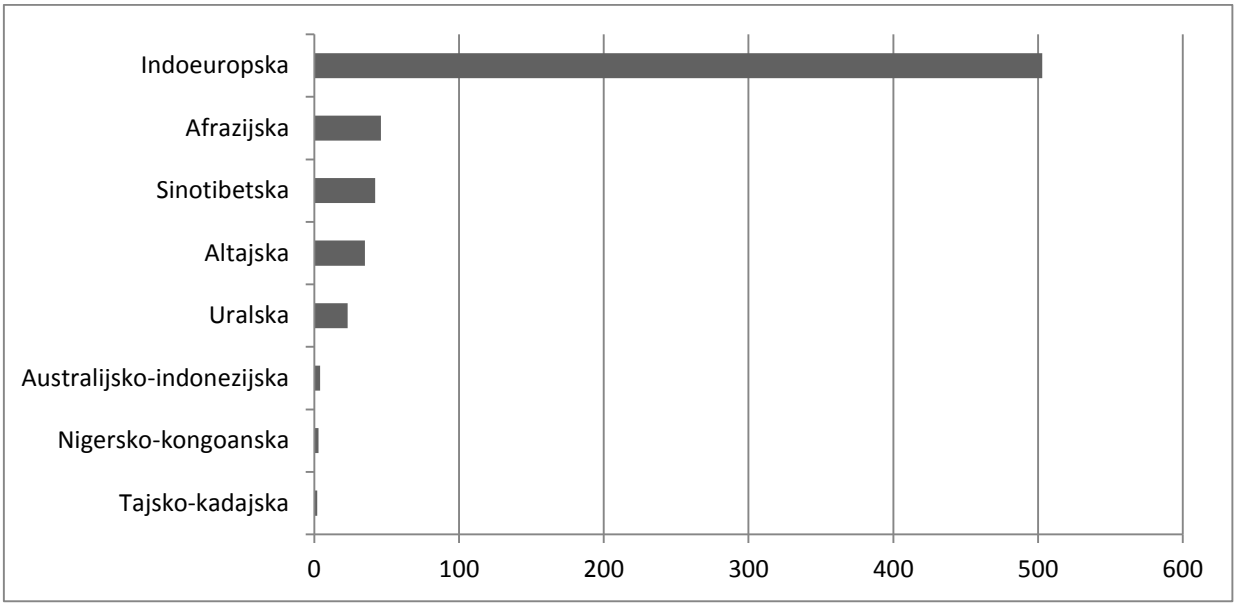
Na sljedećem grafičkom prikazu dan je pregled jezika koji su kao materinski jezici najzastupljeniji među polaznicima *Croaticuma* kroz sva tri analizirana semestra. Najviše je polaznika s španjolskim kao materinskim jezikom, slijede ga engleski, poljski, njemački i kineski. Samo polaznici kojima je materinski neki od ovih pet jezika čine čak 55% svih polaznika u navedenom razdoblju.



Graf br. 12 Materinski jezici s najvećim brojem polaznika

Materinski jezici djelomično otkrivaju porijeklo polaznika te doprinose stvaranju cjelokupne slike o jezičnoj strukturi polaznika. Kroz sva tri semestra daleko najzastupljenija je indoeuropska[202] jezična porodica sa 76% polaznika. Slijede je afrazijska[203], sinotibetska[204], altajska[205] i uralska[206] jezična porodica između kojih postoje male razlike u broju polaznika. Predstavnici afrazijske porodice među polaznicima kao materinski jezici su semitski[207] jezici, a u najvećoj mjeri arapski jezik. Nadalje, sinotibetsku jezičnu porodicu u ovom istraživanju predstavlja kineski jezik, altajsku turski, korejski i japanski te uralsku finski, mađarski i estonski jezik. Preostale skupine jezika s tri različita kontinenta[208] predstavljaju na našim prostorima slabije poznati jezici poput jezika mandinka i luganda[209]. Iako zastupljene u prilično malom broju – zajedno čine svega 2% od ukupnog broja polaznika – znatno doprinose raznolikosti polaznika *Croaticuma*.

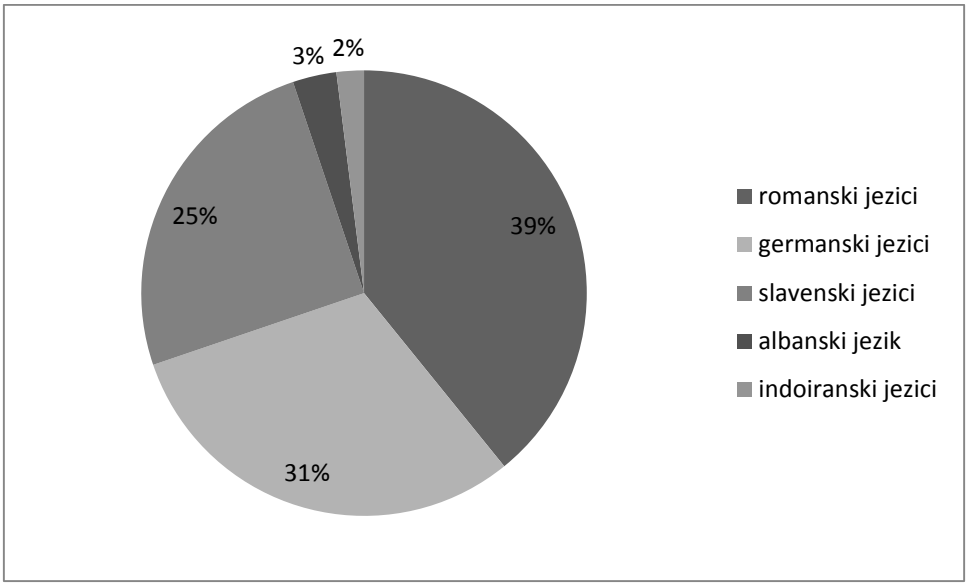
[202] Indoeuropski jezici. // Hrvatska enciklopedija. (4.7.2015.).
[203] Afroasiatic languages. // Wikipedia : the free encyclopedia. (4.7.2015.).
[204] Sino-Tibetan languages. // Wikipedia : the free encyclopedia. (4.7.2015.).
[205] Altaic languages. // Wikipedia : the free encyclopedia. (4.7.2015.).
[206] Uralic languages. // Wikipedia : the free encyclopedia. (4.7.2015.).
[207] Semitic languages. // Wikipedia : the free encyclopedia. (4.7.2015.).
[208] Jezične porodice. Lingvo.info. (4.7.2015.).
[209] Niger–Congo languages. // Wikipedia : the free encyclopedia. (4.7.2015.).



Graf br. 13 Materinski jezici gledani kroz jezične porodice

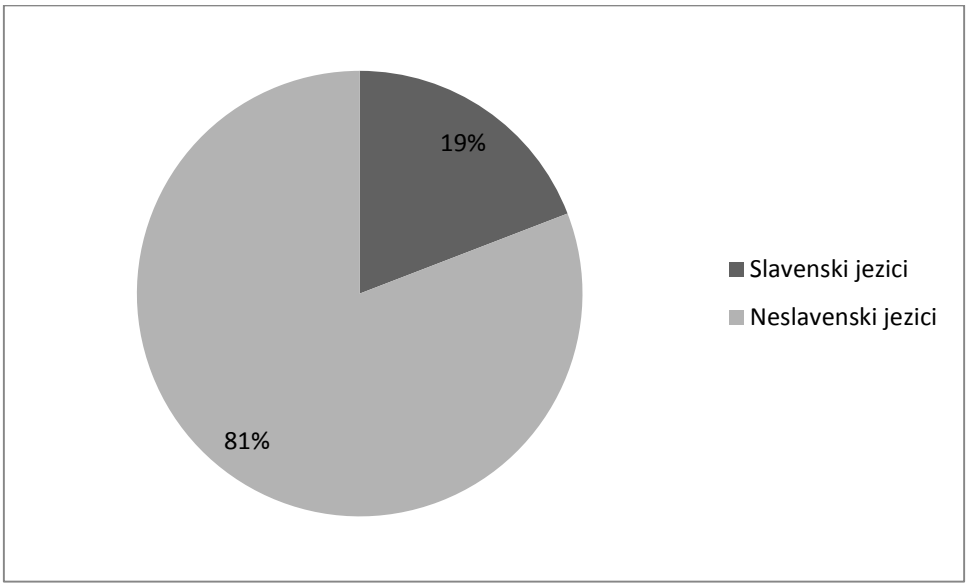
Unutar indoeuropske jezične porodice dominiraju tri najveće skupine jezika: romanska[210], germanska[211] i slavenska[212]. Među njima najzastupljeniji su romanski jezici, uključujući španjolski jezik koji je i u općenitom smislu dominantan među polaznicima *Croaticuma* unutar cijelog promatranog razdoblja. Slijede ih germanski jezici među kojima su kao materinski jezici dominantni engleski i njemački jezik. Osim njih značajni su i sjevernogermanski jezici poput danskog, švedskog i norveškog. Slavenski jezici čine ukupno 25% od ukupnog broja polaznika. U manjem postotku zastupljen je albanski jezik koji čini samostalnu skupinu unutar indoeuropske jezične porodice te indoiranski jezici[213] koje među polaznicima *Croaticuma* predstavljaju perzijski i lari.

[210] Romance languages. // Wikipedia : the free encyclopedia. (4.7.2015.).
[211] Germanic languages. // Wikipedia : the free encyclopedia. (4.7.2015.).
[212] Slavic languages. // Wikipedia : the free encyclopedia. (4.7.2015.).
[213] Indo-Iranian languages. // Wikipedia : the free encyclopedia. (4.7.2015.).



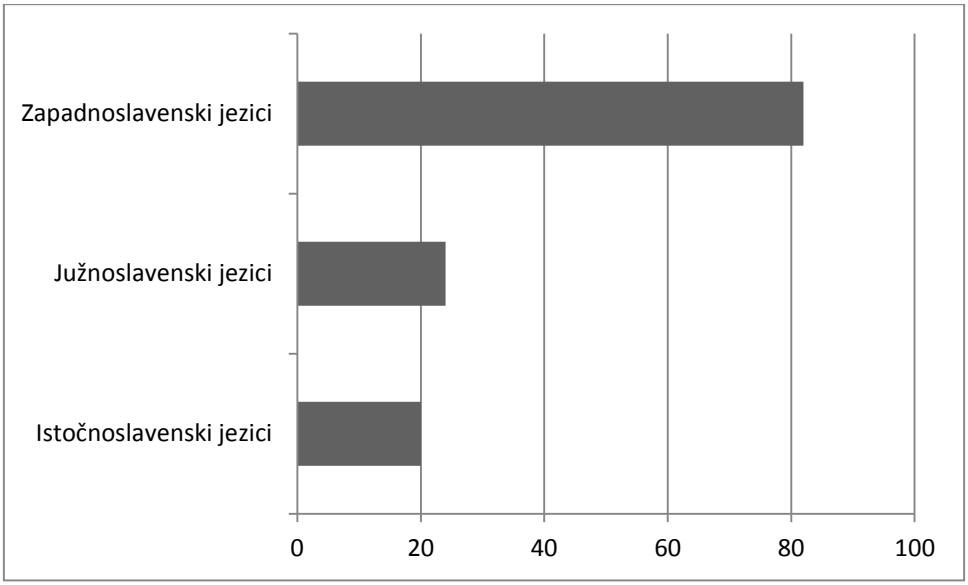
Graf br. 14 Indoeuropska jezična porodica

Ukupno gledano, slavenski jezici u odnosu na sve ostale jezike čine jednu petinu.



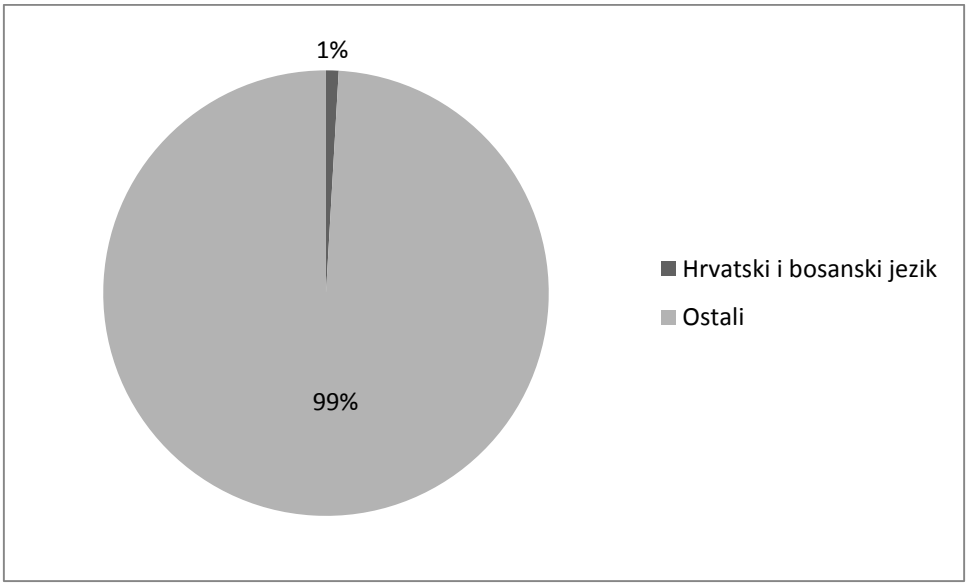
Graf br. 15 Omjer slavenskih i neslavenskih jezika među polaznicima Croaticuma

Unutar slavenske jezične skupine dominiraju zapadnoslavenski jezici predvođeni poljskim koji sam čini 11% od ukupnog broja polaznika i jezika. Istočnoslavenske predstavljaju ruski i ukrajinski s ukupnim udjelom od 16% među slavenskim jezicima. Od južnoslavenskih jezika među polaznicima *Croaticuma* nalazimo slovenski, makedonski, bugarski, hrvatski i bosanski jezik koji zajedno čine 19% od ukupnog broja polaznika među slavenskim jezicima.



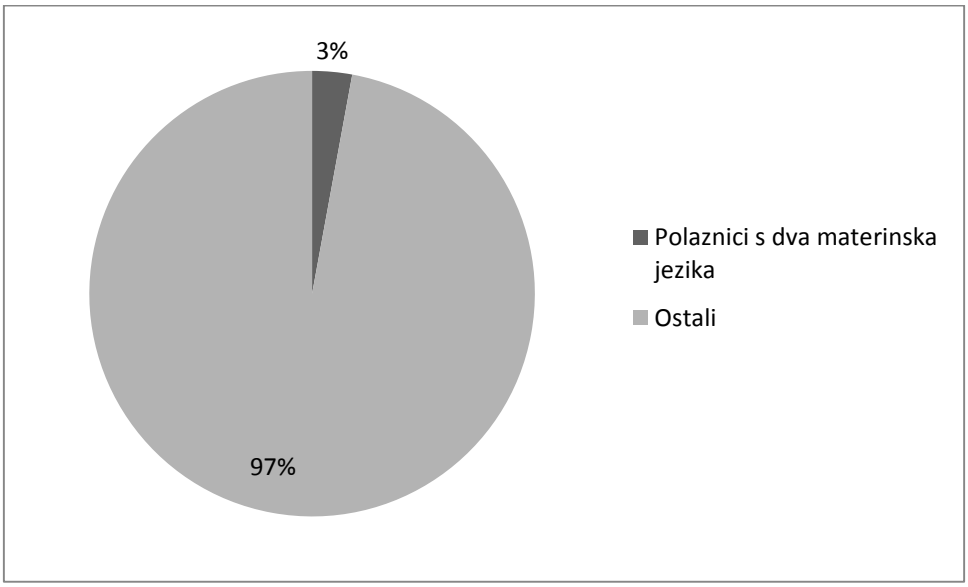
Graf br. 16 Slavenska skupina jezika

Glede južnoslavenskih jezika, hrvatski i bosanski kao materinske jezike, koji dotične polaznike bar u nekoj mjeri čine izvornim govornicima, među polaznicima *Croaticuma* u promatranom razdoblju susrećemo u razmjerno malom broju. Iz analiziranih upitnika nije moguće iščitati razloge zbog kojih ima polaznika koji ove jezike navode kao materinske, a uče hrvatski jezik, no možemo pretpostaviti da je vjerojatno riječ o potomcima iseljenika koji su hrvatski jezik naučili na nekoj razini sporazumijevanja s roditeljima.



Graf br. 17 Omjer hrvatskog i bosanskog jezika i ostalih materinskih jezika

Zanimljiv fenomen predstavljaju polaznici s dva ili čak tri materinska jezika. Uzroci ovog fenomena uglavnom su uvjetovani porijeklom odnosno različitim materinskim jezicima roditelja kao i govornim područjem na kojem se govori jezikom drukčijim od materinskog jezika roditelja. U ovoj skupini među polaznicima *Croaticuma* dominiraju govornici njemačkog jezika.



Graf br. 18 Omjer polaznika s dva materinska jezika i ostalih polaznika Croaticuma

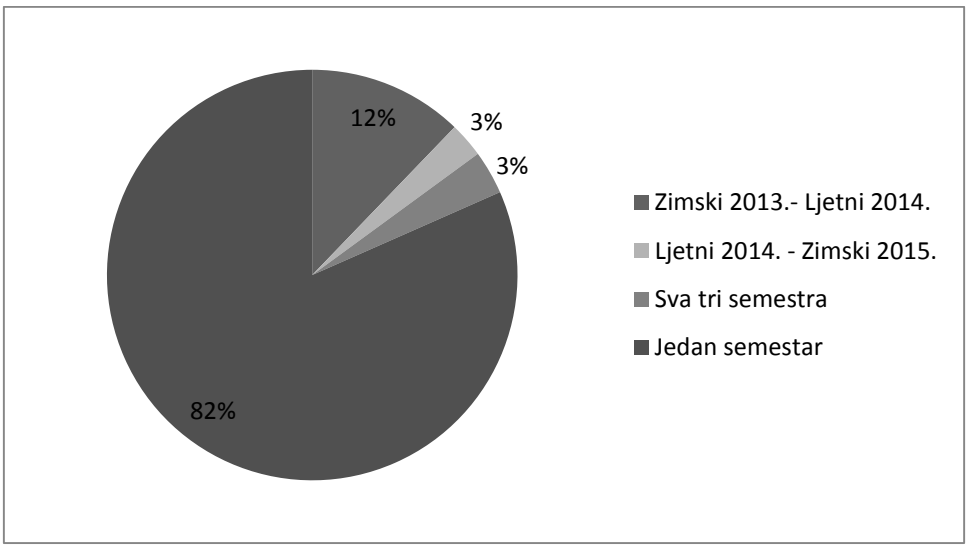
Glede materinskih jezika roditelja polaznika može se reći da su, ako su jednaki kod oba roditelja, u velikoj većini slučajeva jednaki materinskom jeziku djece. A ako su različiti, jedan od njih uglavnom je i materinski jezik polaznika. U nekoliko slučajeva u kojima tomu nije

tako možemo pretpostaviti[214] da je na formiranje materinskog jezika utjecala okolina odnosno govorno područje.

Kad govorimo o godinama učenja hrvatskog jezika, može se reći da uglavnom odgovaraju stupnju učenja na *Croaticumu.*

Duljina boravka u Hrvatskoj nije proporcionalna stupnju učenja hrvatskog, tj. može se reći da rezultati ne slijede nikakvo pravilo. Svaki polaznik je individualan, ima drukčiji jezični razvoj koji pak uvjetuju različite životne okolnosti svakog polaznika. Pritom valja uzeti u obzir da neki polaznici hrvatski uče samo u okviru *Croaticuma,* za vrijeme boravka u Hrvatskoj, dok ga drugi uče i prije polaska na tečaj – bilo u Hrvatskoj bilo u svojoj matičnoj zemlji. Tako, primjerice, među polaznicima na naprednom stupnju učenja hrvatskog nalazimo polaznike koji samo tjedan ili dva borave u Hrvatskoj, a isto tako neke od onih, koji u Hrvatskoj žive već godinu i 6 mjeseci, možemo susresti među polaznicima na početnom stupnju.

Naposljetku, kad govorimo o strukturi polaznika na *Croaticumu*, zanimljivo je promotriti koliki se broj polaznika na *Croaticumu* zadržao dulje od jednog semestra. Dakako, uz podatke za dulje vremensko razdoblje bilo bi moguće dobiti mnogo preciznije rezultate, no unutar ovog istraživanja promatrano je razdoblje od tri semestra. Od ukupnog broja polaznika njih 18% učilo je hrvatski jezik na *Croaticumu* dulje od jednog semestra, a 3% učenika susrećemo u sva tri promatrana semestra.



Graf br. 19 Duljina pohađanja tečajeva na Croaticumu

[214] U upitnicima koje su polaznici ispunjavali nije bilo pitanja o zemlji iz koje potiču ili u kojoj žive.

5.2. Teme učeničkih eseja

Popis pojavnica može se definirati kao lista u kojoj se uz svaku pojavnicu nalazi podatak o njezinoj frekvenciji.[215] U frekvencijskim popisima za korpuse dan je popis pojavnica što znači da svaka riječ čini jedinicu za sebe, tj. prati se frekvencija riječi u svim njezinim oblicima[216], ne samo u osnovnom (npr. nominativ jednine ili infinitiv). Frekvencijski popis unutar ovog rada prilagođen je potrebama istraživanja. Tako su različiti oblici riječi svedeni na osnovni oblik riječi, pridjevi i imenice istog korijena koji upućuju na istu ili sličnu tematiku (npr. obitelj, obiteljski) uzimani su kao jedna pojavnica kao i bliskoznačnice (npr. novac, financije) ili jednostavno riječi sličnog značenja koje s obzirom na kontekst označavaju iste motive (npr. društvene mreže, internet) ili po određenim kriterijima spadaju u istu kategoriju (npr. odjeća, obuća). Izrađen je popis ključnih riječi iz naslova eseja učenika *Croaticuma* za promatrano razdoblje od tri semestra kako bi se stekao uvid u obrađene teme općenito, ali i s obzirom na pojedini stupanj. Može se reći da se kroz tri promatrana semestra uglavnom obrađuju iste teme.

Teme eseja polaznicima se zadaju s obzirom na obrađene nastavne sadržaje koji se pak temelje na općim karakteristikama neke europske zajednice i njezine kulture poput, primjerice, svakodnevice, uvjeta života, odnosa među ljudima, vrijednosti i stavova, običaja i ponašanja u određenim prilikama kako bi učenik, usvajajući jezik s obzirom na navedene aspekte, stekao sposobnost izražavanja u širokom spektru mogućih situacija. Svaka od navedenih tema može se podijeliti na podteme koje predstavljaju pojedine segmente svake od njih:

- svakodnevica: hrana i piće, praznici, posao, slobodno vrijeme
- uvjeti života: životni standard i kvaliteta života
- odnosi među ljudima: odnosi između klasa, spolova, generacija, u obitelji
- vrijednosti i stajališta glede: povijesti, nacionalnih i vjerskih manjina, umjetnosti
- običaji i društvene konvencije prilikom: vjenčanja, vjerskih običaja, različitih proslava itd. [217]

Općenito govoreći, polaznici tečajeva hrvatskog jezika u okviru *Croaticuma* najčešće pišu o sebi, svojoj obitelji, o temama od svog osobnog interesa, o ljudima koje su upoznali, mjestima koja su posjetili, izražavaju svoje dojmove i mišljenje o aktualnim temama. Riječ je

[215] Rječnik korpusne lingvistike. // Hrvatski nacionalni korpus. (4.7.2015.)
[216] Word lists by frequency. // Wikipedia : the free encyclopedia. (4.7.2015.).
[217] Čeliković, V. (ur.) 2005., str 105-106

uglavnom o temama poput zaštite okoliša, hrane i pića, društvenim mrežama i internetu, knjigama, tehnologiji, umjetnosti itd. Česte teme su i Hrvatska i njezini stanovnici – polaznici opisuju svoj doživljaj zemlje i ljudi te ih na temelju različitih aspekata i kriterija uspoređuju sa svojom zemljom. Uvjetno rečeno, rjeđe se piše o složenijim temama, jer takve teme dolaze s naprednim stupnjevima učenja.

Teme su najčešće oblikovane na način da se od polaznika očekuje da nešto opiše – od jednostavnijih poput opisa vlastite sobe, obiteljskog objeda ili osobe ka složenijim poput opisa kvalitete života određenih društvenih skupina. Na naprednim stupnjevima zastupljeni su argumentacijski eseji (prema podacima s *Croaticuma* pišu se počevši od stupnja 1B) koji od polaznika zahtijevaju kritičko promišljanje zadane teme, pogled na situaciju iz različitih kuteva, usporedbu pozitivnih i negativnih strana itd.

moj, moja, moje	44
dom, domovina, rodno mjesto, moja zemlja	19
život, životni	17
obitelj, obiteljski	13
Hrvatska, Hrvati	10
slika	10
opis	9
posao	8
raditi	7
praznici	7
danas, moderan	7
ja	6
dan	6
vrijeme	6
okoliš, ekologija	6
hrana, prehrana, piće	5
vikend	4
kako	4
knjiga	4
jezici	4
novac, financije	4
društvo	4
društvene mreže, internet	4
soba	3
Zagreb	3
Croaticum	3
ljudi	3

odjeća, obuća	3
Božić, božićni	3
mediji	3
običaji	3
veze	3
kvaliteta života, životni standard	3
tipičan	2
kolege	2
film	2
Nova godina, novogodišnja	2
ljubav	2
sreća	2
snovi	2
savršen	2
omiljen	2
komunikacija	2
inozemstvo	2
vjenčanje	2
putovanje	2
volontiranje	2
položaj žene	2
tehnologija	2
umjetnost	2
sport	2
festivali	2
stereotipi	2
rat	2
budućnost	2
svakodnevica	2
uspomene	2
automobili, javni prijevoz	2
restoran	1
pismo	1
reklame	1
razglednica	1
Slavonija	1
Dalmacija	1
snijeg	1
prijateljstvo	1
uskrs	1
vrijednosti	1
izumi	1
ovisnosti	1

aktivizam	1
siromaštvo	1
stariji ljudi	1
ekonomska kriza	1
turizam	1
liječenje	1
religija	1
popularna kultura	1
romantika	1
jednakost	1
Dan žena	1
blagdan Svih svetih	1
sveti Nikola	1
nomen est omen	1
slobodno vrijeme	1
zdravstvo	1
tulumi	1
obrazovanje	1
osobe s posebnim potrebama	1
nacionalne manjine	1

Tablica br. 2 Frekvencijski popis

5.2.1. Karakteristike pismenog izražavanja na pojedinim stupnjevima učenja

Početne stupnjeve 1A i 1A+ karakteriziraju sastavci od najmanje 70 do 100 riječi – prema kraju semestra i nešto više. Najčešće se pišu u obliku domaće zadaće pa su učenici mogli koristiti sva dostupna pomagala poput rječnika i sl. Kao tema često je zadano opisivanje određene slike te eseji na temelju zadanog vokabulara. Prvi stupanj kreće s jednostavnim temama poput obitelji, prijatelja, opisa osoba, odjeće, sobe, pisanja razglednice.

Na stupnju 1B još uvijek prevladavaju prilično jednostavne teme, no ipak nešto složenije nego na prethodnom stupnju. Zadani minimalan broj riječi varira oko 150 do 200. U ovoj fazi učenja polaznici kreću s pisanjem argumentacijskih eseja. Dominiraju teme poput usporedbe običaja ili tradicija između Hrvatske i zemlje polaznika, aktivnosti u slobodno vrijeme te teme koje zahtijevaju izražavanje vlastitih želja i snova.

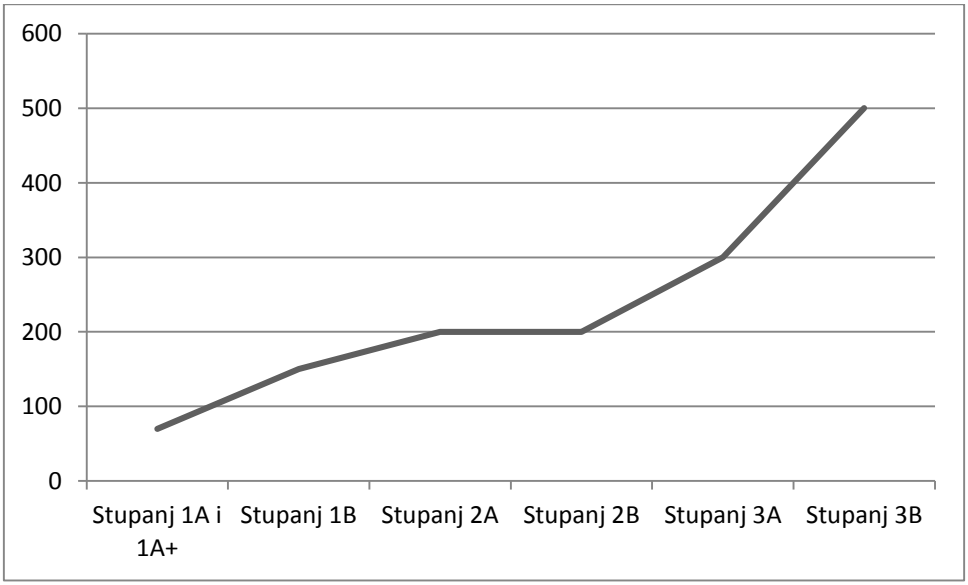
Što se tiče stupnja 2A, može se reći da u odnosu na prethodni stupanj nisu velike razlike u zadanom broju riječi u esejima, no svaki esej je argumentacijski, dakle, cilj je izrada jasne strukture teksta. Struktura eseja mora sadržavati uvod, središnji dio u kojem se razrađuju

prednosti i nedostaci na određenu temu te argumenti kojima se podupiru obje strane.[218] Izbor tema eseja u skladu je s kategorizacijom ovog stupnja kao prijelaznog[219]. Možemo govoriti o svojevrsnoj mješavini koja obuhvaća teme od svakodnevnih (prehrana, vlastita očekivanja) preko osobnih iskustava s nekim općim temama (društvene mreže, umjetnost) do općih tema koje iziskuju i određeno znanje o svijetu, ali i značajan stupanj jezične kompetencije (zaštita okoliša, konzumerizam, moderni izumi).

Što se tiče pismenog izražavanja, stupanj 2B također karakteriziraju argumentacijski eseji koji se uglavnom pišu u okviru domaćih zadataka (i ispita), no teme su mnogo složenije. Iziskuju široko opće znanje i relativno visok nivo jezičnog znanja. Dominiraju teme poput novca, filma, ekologije.

Na naprednim stupnjevima učenja jezika 3A i 3B uz argumentacijski esej kao vrstu teksta susrećemo i interpretacijski esej. Osobitost interpretacijskog eseja jest u tome što polaznik dobiva ulomak iz teksta, pjesmu ili sliku[220] na temelju kojih bi trebao producirati tekst. Očekivana duljina tekstova je od 300 do 500 riječi te više od 500 riječi. Glede tema, naglasak je na kulturološkim osobitostima zemlje poput stereotipa i običaja. Osim toga, zastupljene su i složene teme općenite problematike poput religije, rata, Europske unije i masovnih medija.

Priloženi grafički prikaz prati porast broja riječi u esejima polaznika s obzirom na stupanj, počevši od 70-ak riječi na temeljnom stupnju do više od 500 riječi na naprednim stupnjevima.



Graf br. 20 Rast broja riječi u učeničkim esejima s obzirom na stupanj

[218] Tri vrste školskog eseja. // Hrvatski na mreži. (4.7.2015.).
[219] Čeliković, V. (ur.) 2005., str. 23
[220] Tri vrste školskog eseja. // Hrvatski na mreži. (4.7.2015.).

6. Zaključak

U globaliziranom svijetu današnjice nameće se pitanje je li učenje jezika stvarna potreba ili luksuz. Svjetski jezici, tj. jezici s velikim brojem govornika odavno su se odlučili za prvi odgovor te su i krenuli u tom smjeru izgrađujući resurse za s jedne strane što jednostavnije, a s druge što je moguće učinkovitije učenje stranih jezika. Tako, primjerice, za dominantni jezik današnjice – za engleski jezik, kad govorimo o korpusima u općenitom smislu, valja spomenuti korpus *Bank of English* koji danas obuhvaća desetak potkorpusa s ukupno 650 milijuna riječi te svakodnevno povećava svoj opseg[221].

A razvoj specijalne vrste korpusa u svrhu istraživanja drugog i stranog jezika započinje krajem proteklog stoljeća[222], kad velike nakladničke kuće te stručnjaci s područja učenja stranih jezika postaju svjesni mogućnosti koje pružaju[223]. Nakon objavljivanja korpusa pod nazivom *International Corpus of Learner English* dolazi do pokretanja projekata izrade učeničkih korpusa za takoreći manje popularne jezike u smislu učenja stranih jezika ili jezike s manjim brojem govornika. Brojne su pozitivne strane izrade modernih učeničkih korpusa – počevši od količine građe koju mogu obuhvatiti. Veća količina uglavnom podrazumijeva raznovrsnost građe, a sve zahvaljujući primjeni računala koja zapravo omogućuje samu obradu golemih količina građe.

S težnjom ponovnog usklađivanja sa svjetskim trendovima na ovom području, i za hrvatski je jezik, uz druge slavenske jezike poput ruskog, češkog i slovenskog na Filozofskom fakultetu u Zagrebu pokrenuta izrada prvog hrvatskog učeničkog korpusa kojeg je cilj na temelju autentičnih učeničkih eseja prikupljenih u okviru tečajeva hrvatskog jezika na *Croaticumu,* Centru za hrvatski kao drugi strani jezik Filozofskog fakulteta u Zagrebu stvoriti resurse za izradu različitih aplikacija za interaktivno učenje hrvatskog na različitim stupnjevima jezične kompetencije[224].

[221] The Collins Corpus. // Collins. (4.7.2015.).
[222] Mikelić Preradović; Berać; Boras 2015.
[223] Granger 1998.
[224] Mikelić Preradović; Berać; Boras 2015.

7. Literatura

- About MeLLANGE. // MeLLANGE. 2007. Dostupno na: http://corpus.leeds.ac.uk/mellange/about_mellange.html (4.7.2015.).

- Afroasiatic languages. // Wikipedia : the free encyclopedia. 27.6.2015. Dostupno na: https://en.wikipedia.org/wiki/Afroasiatic_languages (4.7.2015.).

- Altaic languages. // Wikipedia : the free encyclopedia. Dostupno na: 2.7.2015. https://en.wikipedia.org/wiki/Altaic_languages (4.7.2015.).

- Austronesian languages. // Wikipedia : the free encyclopedia. 21.6.2015. Dostupno na: https://en.wikipedia.org/wiki/Austronesian_languages (4.7.2015.).

- Banković-Mandić, I. Izgovorna obilježja učenika hrvatskoga kao drugoga i stranoga jezika na različitim stupnjevima znanja. // Hrvatska znanstvena bibliografija. Dostupno na: http://bib.irb.hr/prikazi-rad?&rad=653576 (18.8.2015.).

- Berman, S. Korpus und Korpuslinguistik. // Ruhr-Universität Bochum. Dostupno na: http://homepage.rub.de/Stephen.Berman/Korpuslinguistik/Allgemeines.html (4.7.2015.).

- Bickel, H. Das Internet als linguistisches Korpus. // Linguistik online. 28, 3 (2006), str. 71-83.

- Bopp, S. Einführung in die Korpuslinguistik mit DeReKo und COSMAS II. // Philologische und historische Fakultät der Universität Augsburg. 9.5.2010. Dostupno na: https://www.philhist.uni-augsburg.de/lehrstuehle/germanistik/sprachwissenschaft/mitarbeiter/stelspass/materialien_lehrveranstaltungen/korpuslinguistik_dereko_cosmas2_bopp.pdf (4.7.2015.).

- Bratanić, M. Korpusna lingvistika ili sretan susret. // Radovi Zavoda za slavensku filologiju. 27 (1992), str. 145-159.

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

- Čeliković, V. (ur.) Zajednički europski referentni okvir za jezike: učenje, poučavanje, vrednovanje. Zagreb : Školska knjiga, 2005.

- Duden Online. // Bibliographisches Institut. Dostupno na: http://www.duden.de/ (4.7.2015.).

- Error analysis. // Wikipedia : the free encyclopedia. 29.4.2015.Dostupno na: http://en.wikipedia.org/wiki/Error_analysis_%28linguistics%29 (4.7.2015.).

- Generativna gramatika. // Hrvatska enciklopedija. Dostupno na: http://www.enciklopedija.hr/Natuknica.aspx?ID=2.1.594 (4.7.2015.).

- Germanic languages. // Wikipedia : the free encyclopedia. 4.7.2015. Dostupno na: https://en.wikipedia.org/wiki/Germanic_languages (4.7.2015.).

- Indo-Iranian languages. // Wikipedia : the free encyclopedia. 27.6.2015. Dostupno na: https://en.wikipedia.org/wiki/Indo-Iranian_languages (4.7.2015.).

- GLBCC. // University of Oxford Text Archive. Dostupno na: http://ota.oucs.ox.ac.uk/headers/2506.xml (4.7.2015.).

- 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. The Learner Corpus: a revolution in applied linguistics. // English Today. 10, 3 (1994), str. 25-33.

- Hansen Kokoruš, R. et al. Njemačko - hrvatski univerzalni rječnik. Zagreb : Nakladni zavod Globus, 2005.

- Hrvatska jezična riznica. // Institut za hrvatski jezik i jezikoslovlje. Dostupno na: http://riznica.ihjj.hr/dokumentacija/index.hr.html (4.7.2015.).

- Hrvatski jezični portal. // Sveučilišni računski centar. Dostupno na: http://hjp.novi-liber.hr/index.php?show=main (4.7.2015.).

- Hrvatski nacionalni korpus. Dostupno na: http://www.hnk.ffzg.hr/default.htm (4.7.2015.).

- ICLE. // Centre for English Corpus Linguistics. 23.6.2015. Dostupno na: http://www.uclouvain.be/en-cecl-icle.html (4.7.2015.).

- Indoeuropski jezici. // Hrvatska enciklopedija. http://www.enciklopedija.hr/natuknica.aspx?ID=27328 (4.7.2015.).

- Indo-European languages. // Wikipedia : the free encyclopedia. Dostupno na: 3.7.2015. https://en.wikipedia.org/wiki/Indo-European_languages (4.7.2015.).

- Indo-Iranian languages. // Wikipedia : the free encyclopedia. 27.6.2015. Dostupno na: https://en.wikipedia.org/wiki/Indo-Iranian_languages (4.7.2015.).

- Jelaska, Z. et al. Hrvatski kao drugi i strani jezik Zagreb : Hrvatska sveučilišna naklada, 2005.

- Jezične porodice. Lingvo.info. Dostupno na: http://lingvo.info/hr/babylon/language_families (4.7.2015.).

- Kapović, M. Uvod u indoeuropsku lingvistiku : pregled jezika i poredbena fonologija. Zagreb : Matica hrvatska, 2008.

- Kilgarriff, A.; Grefenstette, G. Web as Corpus. // Computational Linguistics. 29 (3) (2003). Dostupno na: http://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WACIntro.pdf (18.8.2015.).

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

- Kolipsi : die Südtiroler Schüllerinnen und die Zweitsprache : eine linguistische und sozialpsychologische Untersuchung. // EURAC. Dostupno na: http://www.eurac.edu/de/research/autonomies/commul/projects/Pages/projectdetails.aspx?pid=1818 (4.7.2015.).

- Korpus. // Hrvatski nacionalni korpus. Dostupno na: http://www.hnk.ffzg.hr/korpus.html (4.7.2015.).

- Korpuslinguistik. // Wikipedia : die freie Enzyklopädie. 20.6.2015. Dostupno na: https://de.wikipedia.org/wiki/Korpuslinguistik (4.7.2015.).

- Kosem, I.; Rozman, T.; Stritar, M. How do Slovenian primary and secondary school students write and what their teachers correct: a corpus of student writing. // Proceedings of The Corpus Linguistics Conference. Birmingham, 20. – 22. 7. 2011.

- Kukavica, V. Hrvatski jezik na Internetu : jezične tehnologije za hrvatski. // Zavod za kulturu vojvođanskih Hrvata. Dostupno na: http://www.zkvh.org.rs/batina/jezik/1544-hrvatski-jezik-na-internetu-jezine-tehnologije-za-hrvatski (4.7.2015.).

- Learner corpora around the world. // Centre for English Corpus Linguistics. 23.6.2015. Dostupno na: http://www.uclouvain.be/en-cecl-lcworld.html (4.7.2015.).

- Lemnitzer, L.; Zinsmeister, H. Korpuslinguistik: eine Einführung. Tübingen : Narr Francke Attempto Verlag, 2006.

- Ljubešić, N. Upotreba jezičnih tehnologija u digitalizaciji teksta i njegovoj daljnjoj obradi. // Četvrti festival hrvatskih digitalizacijskih projekata. 10.4.2014. Dostupno na: http://dfest.nsk.hr/2014/wp-content/uploads/2014/04/Nikola_Ljube%C5%A1i%C4%87.pdf (4.7.2015.).

- Macan, Ž.; Kolaković, Z. Prijenosna odstupanja govornika njemačkoga u ovladavanju hrvatskim jezikom. // Lahor. 5 (2008), str. 34-52.

- Mikelić Preradović, N.; Berać, M.; Boras, D. Learner Corpus of Croatian as a Second and Foreign Language. // In: Multidsciplinary Approaches to Multilingualism, Petar Lang, 2015. (u postupku objave)

- Mukherjee, J. Corpus linguistics and language pedagogy: The state of the art – and beyond. // Universität Giessen. Dostupno na: http://www.uni-giessen.de/anglistik/LING/Staff/. mukherjee/pdfs/Mukherjee-2006b.pdf (4.7.2015.).

- Müller, S. Discourse Markers in Native and Non-native English Discourse. // Google Books. 2005. Dostupno na: https://books.google.hr/books?id=WWMpdHUhQ_QC&printsec=frontcover&hl=hr&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false (4.7.2015.).

- National Research University : Higher School of Economics. // Wikipedia : the free encyclopedia. 1.7.2015. Dostupno na: http://en.wikipedia.org/wiki/National_Research_University_%E2%80%93_Higher_School_of_Economics (4.7.2015.).

- Niger–Congo languages. // Wikipedia : the free encyclopedia. 4.7.2015. Dostupno na: https://en.wikipedia.org/wiki/Niger%E2%80%93Congo_languages (4.7.2015.).

- Pranjković, I. Jezikoslovlje i tekst. // Vijenac. 401 (2009). Dostupno na: http://www.matica.hr/vijenac/401/Jezikoslovlje%20i%20tekst/ (4.7.2015.).

- Rječnik korpusne lingvistike. // Hrvatski nacionalni korpus. Dostupno na: www.hnk.ffzg.hr/bb/definicijekl.doc (4.7.2015.).

- Romance languages. // Wikipedia : the free encyclopedia. 21.6.2015. Dostupno na: https://en.wikipedia.org/wiki/Romance_languages (4.7.2015.).

- RULEC : Russian Learner Corpus of Academic Writing. // Web corpora. Dostupno na: http://web-corpora.net/RussianLearnerCorpus/search/ (4.7.2015.).

- Second language acquisition. // Wikipedia : the free encyclopedia. 21.5.2015. Dostupno na: http://en.wikipedia.org/wiki/Second-language_acquisition (4.7.2015.).

- Second language. // Wikipedia : the free encyclopedia. Dostupno na: https://en.wikipedia.org/wiki/Second_language (4.7.2015.).

- Semitic languages. // Wikipedia : the free encyclopedia. 29.6.2015. Dostupno na: https://en.wikipedia.org/wiki/Semitic_languages (4.7.2015.).

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

- Sino-Tibetan languages. // Wikipedia : the free encyclopedia. 22.6.2015. Dostupno na: https://en.wikipedia.org/wiki/Sino-Tibetan_languages (4.7.2015.).

- Skraćeni semestralni tečaj. // Croaticum. Dostupno na: http://croaticum.ffzg.unizg.hr/?page_id=856&lang=hr (4.7.2015.).

- Slavic languages. // Wikipedia : the free encyclopedia. 31.5.2015. Dostupno na: https://en.wikipedia.org/wiki/Slavic_languages (4.7.2015.).

- Specijalizirani intenzivni tečaj. // Croaticum. Dostupno na: http://croaticum.ffzg.unizg.hr/?page_id=858&lang=hr (4.7.2015.).

- Speech corpus. // Wikipedia : the free encyclopedia. 20.1.2015.Dostupno na: https://en.wikipedia.org/wiki/Speech_corpus (4.7.2015.).

- Srpskohrvatski jezik. // Wikipedija : slobodna enciklopedija. 30.6.2015. Dostupno na: http://hr.wikipedia.org/wiki/Srpskohrvatski_jezik (4.7.2015.).

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

- Südtirol. // Wikipedia : die freie Enzyklopädie. 2.7.2015. Dostupno na: https://de.wikipedia.org/wiki/S%C3%BCdtirol (4.7.2015.).

- Šojat, K.; Srebačić, M.; Štefanec, V. CroDeriV i morfološka raščlamba hrvatskoga glagola. // Suvremena lingvistika. 75 (2013), str. 75-96.

- Tadić, M. Building the Croatian National Corpus. // Proceedings of the Third International Language Resources and Evaluation Conference. Pariz, 2002. Str. 441-446.

- Tadić, M. Jezične tehnologije i hrvatski jezik. Zagreb : Ex libris, 2003.

- Tadić, M. Od korpusa do čestotnog rječnika hrvatskoga književnog jezika. // Radovi Zavoda za slavensku filologiju. 27 (1992), 169-178.

- Tadić, M. Računalna obradba hrvatskih korpusa povijest, stanje i perspektive. // Suvremena lingvistika. 43/44 (1997), str. 387-394.

- Tadić, M. Računalna obradba hrvatskoga i nacionalni korpus. // Suvremena lingvistika. 41/42 (1996), str. 603-611.

- Tadić, M. Raspon, opseg i sastav korpusa hrvatskoga suvremenog jezika. // Filologija. 30 – 31 (1998), str. 337-347.

- Tai–Kadai languages. // Wikipedia : the free encyclopedia. 20.6.2015. Dostupno na: https://en.wikipedia.org/wiki/Tai%E2%80%93Kadai_languages (4.7.2015.).

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

- Textkorpus. // Wikipedia : die freie Enzyklopädie. 20.6.2015. Dostupno na: https://de.wikipedia.org/wiki/Textkorpus (4.7.2015.).

- The Collins Corpus. // Collins. Dostupno na: http://www.collins.co.uk/page/The+Collins+Corpus (4.7.2015.).

- The computer learner corpus: A versatile new source of data for SLA research. // Learner English on Computer / Sylviane Granger. London : New York : Addison Wesley Longman, 1998. Str. 3-18.

- The CzeSL-plain corpus. // Institute of the Czech National Corpus. Dostupno na: https://ucnk.ff.cuni.cz/english/czesl-plain.php (4.7.2015.).

- The MeLLANGE Learner Translator Corpus. // MeLLANGE. 2007. Dostupno na: http://corpus.leeds.ac.uk/mellange/ltc.html (4.7.2015.).

- Tri vrste školskog eseja. // Hrvatski na mreži. Dostupno na: http://sikavica.joler.eu/drzavna-matura/eseji/upute-za-pisanje-eseja (4.7.2015.).

- Uralic languages. // Wikipedia : the free encyclopedia. 16.6.2015. Dostupno na: https://en.wikipedia.org/wiki/Uralic_languages (4.7.2015.).

- Više djevojaka upisuje studij i više ih završava. // Hrvatska radiotelevizija. Dostupno na: http://vijesti.hrt.hr/116846/vise-djevojaka-upisuje-studij-i-vise-ih-zavrsava (5.9.2015.)

- Word lists by frequency. // Wikipedia : the free encyclopedia. 10.6.2015. Dostupno na: https://en.wikipedia.org/wiki/Word_lists_by_frequency (4.7.2015.).