

Zer i(ra)kas dezakegu geure corpusekin "jolastuz"?

Mikel Iruskietza
UPV/EHUko IXA taldea
Hizkuntza eta Literaturaren
Didaktika Saila
mikel.iruskietza@ehu.eus

Arantxa Otegi
UPV/EHUko IXA taldea
Lengoaia eta Sistema
Informatikoak Saila
arantxa.otegi@ehu.eus

Larraitx Urria
UPV/EHUko IXA taldea
Lengoaia eta Sistema
Informatikoak Saila
larraitx.urria@ehu.eus

Arantza Diaz de Ilaraza
UPV/EHUko IXA taldea
Lengoaia eta Sistema
Informatikoak Saila
a.diazdeilaraza@ehu.eus

Amaia Artolazabal
Euskal Herriko Ikastolen
Elkartea
Didaktika saila
aartolazabal@ehi.ikastola.eus

Laburpena

Hizkuntzak ikasteko askotariko metodologiak erabili izan dira: metodo zuzena, itzulpen metodoa, metodo audiolinguala, metodo komunikatiboa, [hurbilpen lexikoa](#), ariketetan oinarritutako metodoa, ikasleen erroreetan oinarritutakoa edota metodo eklektikoak. Azken urteotan, berriz, corpusekin "jolasteak" hizkuntzak modu esanguratsuan i(ra)kasteko aukerak eskaintzen dizkigunaren ustean gaude. Corpusekin "jolastuz" ikasteko, ezinbestekoa da tresna informatiko egokiak eskura izatea, zailtasun teknologiko handirik aurkezten ez duten eta edonork modu errazean balia ditzakeen tresnak prestatzea. Hizkuntzaren hainbat aspektu, egitura, erabilera edota ezaugarri ezagutzeko, aztertze eta prozesatzeko, Hizkuntzaren Azterketa eta Prozesamenduaren (HAP) alorrean askotariko teknologiak edota baliabideak gara daitezke; besteak beste, testu-analisiak egiten dituzten tresnak eta analizatutako testuak kontsultatzeko baliabideak. Artikulu honen helburua da, hain zuzen ere, norberari interesatzen zaizkion testu-bildumak hizkuntza-teknologiekin aztertze dauden aukera batzuk aurkeztea eta lortutako datuekin "jolastuz" hizkuntzen i(ra)kaskuntzaren prozesuan atera daitezkeen hainbat ondorio nahiz egin daitezkeen hainbat ekarpen planteatzea, ikasleak bere ikaskuntza planifikatzeko eta norbere ikaskuntzaren kontzientzia hartzeko. Proposamen hau CLARIN-K proiektuaren barruan kokatzen da, zeinaren xede nagusia erabiltzaileentzat tresna erabilerrazak garatzea eta horiek erabiltzeko aholkularitza eskaintzea den.

Hitz gakoak: Humanitate Digitalak, hizkuntzen irakaskuntza, corpusak, teknologia.

Abstract

Different approaches have been used as language learning techniques: direct methods, translation based methods, audiolingual methods, communicative methods, [lexical approach](#), task-based approaches, methods based on learner errors or eclectic methods. In recent years, it is thought that "playing" with a corpus is an interesting way to learn it. However, in order to "play" with a corpus, it is essential to have available appropriate language technologies, resources that do not present great technological difficulties and can be easily used by anyone. Different aspects, structures, uses or characteristics of a language can be analysed and processed using language technologies developed in the field of the Natural Language Processing (NLP), tools for text analysis and applications for searching the analysed data, among others. The main goal of this article is to present some language technologies that offer users the possibility to analyse their own texts and the contributions one can make "playing" with the analysed corpora, to plan their learning and to be conscious of the learning process. This proposal is a task of the CLARIN-K project, whose main objective is to develop accessible tools and provide advice on how to use them to every user.

Keywords: Digital Humanities, language teaching, corpora, technology.

1. SARRERA

Gaur egun, jada, gero eta corpus handiagoak eta askotarikoagoak ditugu eskura euskarri elektronikoa, eta corpusen erabilera ere nabarmen handitu da azken urteotan. Hala, corpusekin “jolastea” hizkuntza bat i(ra)kasteko metodo osagarri bat izan daiteke (Gabrielatos, 2005). Izan ere, tresna informatiko baten laguntzarekin, hizkuntzaren hainbat aspektu, egitura, erabilera edota ezaugarri ezagutzeko aukerak eskaintzen dizkigute corpusek, eta analisi horien baitan azter daiteke testuetan erabili den hizkuntza, planifika daitezke ikas-metodologiak edota sortu ikas-material berriak, adibidez.

Ingelesaren kasuan, esaterako, 1987. urte inguruan hasi ziren ingeleseko irakasleak corpusei arreta jartzen, corpusetan oinarritutako ikasleentzako lehenengo hiztegia (*Collins COBUILD English Language Dictionary*) argitaratu zenean. Urtebete beranduago, gai honen inguruan oihartzun handia izan zuen artikulu bat argitaratu zen, corpusen araberrako edota corpusetan oinarritutako materialak hizkuntzen irakaskuntzan erabiltzeari buruzkoa, *Whence and whither classroom concordancing* (Johns, 1988).

Orduz geroztik, hizkuntzaren egitura eta erabilera ikertzeko corpusetan oinarritzen diren hainbat lan argitaratu dira (McEnery eta Hardie, 2011, esaterako), eta horrekin batera, corpusetan oinarritutako hizkuntzen i(ra)kaskuntzaren nahiz material didaktikoen prestaketaren inguruko lanak ere ugaritu dira (Meunier, 2011; Leech, 2011; Granger ~~eta~~ Meunier, 2010). Eta oso pixkanaka bada ere, hizkuntza irakasleak nahiz ikasleak corpusetan oinarritutako materialen (hiztegiak eta gramatikak, batik bat) kontsumitzaile gero eta ohikoagoak bihurtzen ari dira.

Euskararen kasuan, IXA taldean, esaterako, hainbat proiektu burutu dira corpusek eta Hizkuntzaren Azterketa eta Prozesamenduko (HAP) tresnek hizkuntzen i(ra)kaskuntzan izan dezaketen eraginaren inguruan eta ordenagailuek ikasgelan eta ikasgelatik kanpo hizkuntzak i(ra)kasteko eskaintzen dituzten aukeren inguruan: “hizkuntzen irakaskuntzaren arloan aukera asko ematen ditu informatikak” (Uria ~~et al.~~, ~~Maritxalar eta Zabala~~, 2014; Aldabe ~~et al.~~, ~~Maritxalar eta Lopez de Lacalle~~, 2013; Aldabe ~~et al.~~, 2012; Uria ~~et al.~~, ~~Maritxalar eta Aldabe~~, 2011; Uria ~~et al.~~, ~~Maritxalar eta Zabala~~, 2010; Aldabe ~~et al.~~, 2008). Eta hizkuntzen i(ra)kaskuntzarako garatu izan diren HAPeko tresna asko corpusetan oinarritutakoak dira, bai corpus orokorretan eta baita ikasleek beraiek sortutako corpusetan (alegia, ikasleek idazten dituzten testu bildumetan).

Corpusekin lan egin ahal izateak autonomia ematen dio ikasleari eta metodologia horri jarraiki, ikasle nahiz irakaslearen rola aldatu egiten dira: gramatika arauak zeintzuk diren azaldu beharrean, ikasleari corpusak osatzen eta aztertzen laguntzen dio irakasleak, corpusetan dauden gramatika arauak aurkitzen eta egitura jakinak lantzen; horrela, ikasleak ikaskuntza esanguratsuago eta kontzienteago bat lortzen du. Beraz, batetik, ikasleari bere kabuz ikasteko aukerak eskaintzen zaizkio, eta bestetik, irakaslearen lana ere aldatzen da eskura baliabide berriak izanik (ez du testu-liburuetara soilik mugatu beharrik, baizik eta corpusek bestelako hainbat aukera eskaintzen dizkiote, bai ikaslearen ikasketa-prozesua jarraitzeko eta baita corpusetako informazioa kontuan izanik material edota ariketa berriak prestatzeko).

I(r)akasleen artean, ordea, corpusen erabilera eta hizkuntza-teknologiak nahiko ezezagunak direla iruditzen zaigu, nahiz eta hizkuntza aztertzeko, lantzeko edota ikasteko eskaintzen dizkiguten aukerak handiak diren. Irakasle askok galdetzen digute zer den corpus bat, zeintzuk diren hizkuntzen i(ra)kaskuntzarako egokienak izan daitezkeen corpusak edota nola erabil daitezkeen corpusak hizkuntzak i(ra)kasteko. Iruditzen zaigu, halaber, irakasle askok oraindik ere ez dutela baliabide berriak beraien eskoletan sartzeko behar beste azalpenik edo laguntzarik jaso. Horregatik,

artikulu honetan corpusek eta hizkuntza-teknologiek eskain ditzaketen hainbat aukera aurkeztu nahi ditugu, tresna horiek ezagutarazteko eta teknologia horiekin testu-bilduma handien azterketak egiteko dauden aukeren berri emateko.

Corpusen analisirako hainbat baliabide garatu izan dira beste hizkuntza batzuetan. *Lextutor* webgunea¹, esaterako, oso baliagarria da ikertzaile nahiz irakasleentzat, hainbat hizkuntzatan (ingelesez, frantsesez, gaztelaniaz, alemanez) gramatika eta hiztegia i(ra)kasko baliabide interesgarriak biltzen dituelako (hala nola, corpusak garbitzeko tresnak: puntuazioa, irudiak eta karaktere arraroak kentzeko programak, erro bera duten hitzak erauzteko programa, hitz-zerrendak sortzeko programak, hutsuneak betetzeko ariketak egiteko tresnak, irakurtzen ari zarela hitzak entzuteko, hitz batean klik eginez konkordantziak ikusteko eta esanahia ulertzeko tresnak), eta ikasleak berak bere kabuz ikasko moduko ariketak egin ditzakeelako (hiztegia eta gramatika arauen kontsulta, besteak beste). Webgune horietan, hitzak eta hitz horien testuinguruak aztertzekeo tresnak daude (konkordantziak), hitzen esangura eta erabilera testuinguru anitzetan erabiltzen ikasko. Horrez gain, askotariko ariketak sor daitezke eskolarako, ariketa mota horiek ebaluatuak izan direlarik ikasleekin testuinguru errealean.

Horrez gain, badaude testu analisirako beste tresna automatiko batzuk ere. *Coh-Metrix*² baliagarria da, adibidez, testuen kohesioa eta koherentzia-neurriak konputatzeko, edo *Text Inspector*³, ingelesezko testuen zailtasun-maila aztertzekeo tresna da, maila jakina eta deskribapen interesgarriak ematen dituelarik.

Bide horretan lehen pausoa UPV/EHuko IXA ikerketa-taldean eman dugu eta testu-analisirako tresna bat garatu da: ANALHITZA, zeinak oso modu erraz eta bizkorrean testu bati dagokion informazio linguistikoa osatua ematen dion erabiltzaileari. CLARIN-ERIC europar azpiegituraren barruan kokatzen den K-zentro batean garatu da ANALHITZA, IXA-CLARIN K-zentroan, hain zuzen ere (Bel *et al.*, 2016), zeinaren asmoa den Hizkuntzaren Azterketa eta Prozesamenduko (HAP) tresnak garatzea, erabilgarri jartzea eta horiek erabiltzeko aholkularitza eskaintzea Euskal Herriko irakasle eta ikertzaileei. Horrela, honako lan hau azken urteotan interes handia piztu duen Humanitate Digitalen alorrean kokatzen da.⁴

Hala, laburtuz, artikulu honen asmoa da testuak aztertuz eta analisi horiekin “jolastuz” hizkuntzen i(ra)kaskuntzarako lor daitezkeen emaitzak eta baliabideak aurkeztea.

Artikuluaren bigarren atalean corpusei buruz mintzatuko gara. Hirugarren atalean, corpusek hizkuntzen i(ra)kaskuntzan nola lagun dezaketen azalduko dugu. Laugarren atalean, norbere testuak aztertzekeo garatu diren ANALHITZA eta Voyant Tools tresnak aurkeztuko ditugu. Bosgarren atalean, ANALHITZArekin egin ditugun analisi edota esperimentu batzuetatik ateratako ideia garrantzitsuenez, adibideez eta ondorioez hitz egingo dugu. Azkenik, seigarren atalean, atera ditugun ondorioak eta etorkizuneko lanak aipatuko ditugu.

1 <https://www.lexutor.ca/>

2 <http://tool.cohmetrix.com/>

3 <http://www.textinspector.com/>

4 Humanitate Digitalen alorraren xedea baliabide informatikoak humanitateen barruan kokatzen diren diziplinetan (filologia, filosofia, soziologia, historia, artea, gizarte-zientziak, liburutegiak, etab.) txertatzea da, ikerkuntza hedatzeko eta hobetzeko.

2. CORPUSAK, INFORMAZIO ITURRI ABERATSAK ETA FIDAGARRIAK

Corpusa irizpide zehatz batzuen arabera egituratuta dagoen testu-multzo elektronikoa da, erabilera errealak jasotzen dituen eta, beraz, hizkuntzaren atal edota bariazio baten erakusgarri gisa erabili ohi dena (ikus Aldezabal *et al.* (2005) corpusei buruzko informazio gehiagorako). Irizpide jakin batzuk jarraituz osatu eta antolatu behar da corpusa, ikerketarako baliagarria izango den lagin adierazgarri bat izatea nahi badugu. Eta corpus adierazgarriak, oro har, hizkuntzaren erabilera errealak eta fidagarriak dira; literatur testuek edota prentsako testuek, esaterako, idatzizko hizkuntza eredugarria, zuzena eta formala islatu ohi dute; ikasleen testuek, aldiz, egin ohi dituzten erroreak edota erabilera islatzen dituzte, besteak beste.

Hiztegietan ez bezala, corpusetan hitzak testu barruan agertzen dira, beren testuinguruarekin eta komunikazio-asmoarekin. Horregatik, corpusak baliagarriak dira oso, hitzei eta egiturei buruzko informazioa eskuratzeko, hitz edota egitura jakin bat testuan nola erabili den jakiteko (zein testu motatan, zein testuingurutan, zein egituratan, zein hitzekin batera, zein adierarekin, zeren itzulpen gisa, besteak beste). Esate baterako, egitura baten erabilera zuzenaren inguruan zalantzarik badugu, edo zalantzazkoak zaizkigun bi egituratatik egokiena zein den ez badakigu, corpusetara jo eta bertan ikus dezakegu nola erabili izan den egitura hori (denbora gutxian erabilitako adibide asko jasoz). Adibide gisa, demagun ez dakigula zein den forma zuzenagoa: *batean baino gehiagotan* ala *bat baino gehiagotan*. Egitura jakin horiek bilatu daitezke corpus digitaletan. Ereduzko Prosa Dinamikoa (EPD) corpusean, esaterako, *batean baino gehiagotan* egituraren 68 agerraldi daude (1. irudia) eta *bat baino gehiagotan* egitura, berriz, 30 aldiz agertzen da (beti ere adibideen iturburua zehatzuz). Beraz, corpusean bilatutako egituraren maiztasuna kontuan izanik, lehenengoaren alde egingo genuke.

1. irudia: *batean baino gehiagotan* egituraren bilaketaren adibidea eta emaitzak, EPD corpusean.

Ereduzko Prosa Dinamikoa (EPD)

Besteren berri: Egungo Euskararen Hiztegia 'b' letra osorik argitaratu da (2017-01-30)

Corpus arakatzailerak | Maiztasunak

Liburuak / Prentsa
Bilatu Hustu
lehen urte bitartea: 2011-15
bigarren urte bitartea: 2004-08

1. hitza: batean
Morfologia:
distantzia: 1

2. hitza: baino
Morfologia:
distantzia: 1

3. hitza: gehiagotan
Morfologia:
distantzia: 1

4. lema:
Morfologia:
distantzia: 1

5. lema:
Morfologia:
distantzia: 1

2011-15 2004-08

Orokorra Prentsa Liburuak Corpusean murriztu Autoreak

Emaitza: 63 agerraldi / 21 esaldi
Liburuetan: 9 esaldi / 8 liburu
Prentsan: 12 esaldi / 12 artikulua

Kirola, BERRIA	12 agerraldi / 4 esaldi
Kultura, BERRIA	12 agerraldi / 4 esaldi
Mundua, BERRIA	6 agerraldi / 2 esaldi
Euskal Herria, BERRIA	6 agerraldi / 2 esaldi
Iazko hezurak, UNAI ELORRIAGA	6 agerraldi / 2 esaldi
Biodiskografiak, IBAN ZALDUA	3 agerraldi / esaldi 1
Lili eta biok, RAMON SAIZARBITORIA	3 agerraldi / esaldi 1
Laranja bat zaborretan, PATXI ZUBIZARRETA	3 agerraldi / esaldi 1
Ulisea, JAMES JOYCE / XABIER OLARRA	3 agerraldi / esaldi 1
Martutene, RAMON SAIZARBITORIA	3 agerraldi / esaldi 1
Kartografiak eta mikrobio narratiboen artean, AMAIA ITURBIDE	3 agerraldi / esaldi 1
Benito Cereno / Billy Budd, marinela, HERMAN MELVILLE / JOSE RAMON VAZQUEZ	3 agerraldi / esaldi 1

1. irudia: *batean baino gehiagotan* egituraren bilaketaren adibidea eta emaitzak, EPD corpusean.

Kontuan izan behar da, hala ere, hitz edota egitura baten agerpen kopuruak ez duela bermatzen beti hitz edota egitura horren erabilera egokiena. Alegia, aintzat hartu behar da corpus batzuk —nolabait esatearren— testu-biltegi hutsak izan daitezkeela, inolako iragazkirik edota iturriari buruzko argibiderik gabeak eta, beraz, testu horietatik jasotzen den informazioa tentuz baliatu beharko dela.

Horregatik, garrantzitsua da helburuen arabera corpus bat ala beste aukeratzea. Corpora lagin bat izango da, aztertu nahi den hizkuntzaren lagin bat (ikasleen corpora, corpus dialektal bat edota corpus historiko bat), eta garrantzitsua da lagin hori ahalik eta eredugarriena izatea, emaitza adierazgarriak atera ahal izateko.

Bestalde, bi metodologia bereizten dira corpusekin lan egitean:

1. Corpusetan oinarritutako analisiak (*Corpus Based Analysis* ingelesez): intuizioan du oinarria eta corpusetan egiaztatutako ikerketek corpusak baliatzen dituzte ikerlariaren teoriak edota hipotesiak betetzen ote diren aztertzeko, hipotesi horiek baliozkotzeko, ezeztatzeko edota egokitzeko helburuarekin. Honetan datza, hain zuzen ere, corpusetan oinarritutako hizkuntzalaritza (*corpus linguistics*).
2. Corpusek gidatutako analisiak (*Corpus Driven Analysis* ingelesez): corpusak berak izan behar du gure hizkuntza-hipotesien edota teorien iturri bakarra, ~~alegia, corpora-bera da hizkuntzaren teoriaren oinarri bakarra, hau~~ hau da, corpora izango da hipotesi edota aurreuposiziorik gabe, marko teorikoa zehazteko, datuak ateratzeko eta fenomeno linguistiko jakinak detektatzeko oinarri enpirikoa. Beraz, corpusek gidaturiko hizkuntzalaritzak baztertu egiten du intuiziotik abiatu eta corpusetan intuizio horiek baieztatzeko metodoa.

Euskarazko corpus digitalei dagokienez, askotariko corpus motak daude kontsultagarri. Corpus elebakarrak dira hizkuntza bakar bateko testuez osatzen direnak, eta izan daitezke orotariko corpusak (egunkari, aldizkari eta eleberrietatik hartutako testuez osatuak) edota corpus espezializatuak (gai jakinei edo teknikoei buruzko testuez osatuak); corpus elebidunak edota eleaniztunak, izan daitezkeenak paraleloak edo ez-paraleloak; ahozko corpusak edota idatzizkoak; baliabide interesgarriak guztiak, ezbairik gabe, hizkuntzaren hainbat aspektu aztertzeko abiapuntu gisa. Batzuk aipatzearen, UPV/EHUko webgunean⁵ ondoko corpus hauek daude kontsultagarri: Egungo Testuen Corpora (ETC), Ereduzko Prosa Gaur (EPG), Ereduzko Prosa Dinamikoa (EPD), Euskal Klasikoen Corpora (EKC), Hizkuntzen Arteko Corpora (HAC), Goenkale Corpora eta Pentsamendu Klasikoak. Eta euskarazko beste corpus aipagarri batzuk dira XX. mendeko Euskararen Corpus Estatistikoa⁶, Corpeus⁷ edota Zientzia eta Teknologia Corpora⁸, baita Linguee⁹ corpus eleanitza ere, batzuk aipatzearen.

Artikulu honetan aurkezten ditugun tresnek ere corpusetan bilaketak egiteko aukerak eskaintzen dituzte, baina norberari interesatzen zaizkion testuak dira aztertzen direnak, eta ez aipatu ditugun moduko corpus orokor edota finkoak.

Corpusak eskuz aztertzeara, baina, lan astuna da, nekagarria (eta tamainaren arabera, ezinezkoa). Horregatik, corpusen eskuzko azterketak lan berezietan bakarrik egiten dira, denbora eta tamaina mugatuan azter daitezkeen corpusekin. Egun, Humanitate Digitalen alorrean, ordenagailuz lagundutako metodoak nagusitzen ari dira, eta horrek eragin nabarmena izan du corpusekin lan egiteko moduetan: corpusak biltzeko, egituratzeko eta analizatzeko beharra sortu da, eta horrek baliabide teknologiko berrien sorrera ekarri du. Hala, gero eta tresna gehiago ditugu eskura, corpusetako informazio linguistikoa modu erraz eta sinplean eskuratzeko, testuak analizatu eta informazio horrekin jolasteko aukera ematen digutenak.

5 <http://www.ehu.es/eu/web/eins/egungo-testuen-corpora-etc->

6 <http://xxmendea.euskaltzaindia.eus/Corpus/>

7 <http://corpeus.elhuyar.eus>

8 <http://www.ztcorpusa.eus>

9 <http://www.linguee.es/>

3. CORPUSAK HIZKUNTZEN I(RA)KASKUNTZAN

Esan bezala, beraz, hizkuntzen i(r)a kaskuntzaren alorrean ere, egun, jada, gero eta gehiago erabiltzen dira corpusak bitarteko edo material didaktiko gisa. Erabili ohi diren corpusak orotariko erreferentzia corpusak, ikasleekin lantzeko moduko testuak (adinera edota hizkuntza-mailara egokituak) eta ikasleek beraiek sortutako testuak¹⁰ izan ohi dira.

Ikasleen corpusak oso interesgarriak dira, bai irakasleentzat, bai ikasleentzat eta baita ikerlariontzat ere. Irakasle batek, bere ikasleen testuak aztertuz, ikus dezake zeintzuk diren ikasleek askotan egiten dituzten errore-motak, zein eremutan dabilzan ongi eta zeinetan behar duten laguntza. Eta corpusak aztertzen irakatsiz gero, ikasleek beraiek, testuak aztertuz eta konparatuz, landu ahalko lituzkete zailtasunak ematen dizkieten egitura horiek, bakoitzak bereak, horrela, gainera, beraien ikaskuntza-prozesuan ardurak hartzen ere ikasten dutelarik. Ikerlariok ere, gure aldetik, corpus horiekin askotariko ekarpenak egin ditzakegu alor honetan, azterketa horien baitan hizkuntzak i(r)a kaskuntzako metodologiak nahiz ikas-material berriak prestatzeko aukera berriak sortzen direlako.

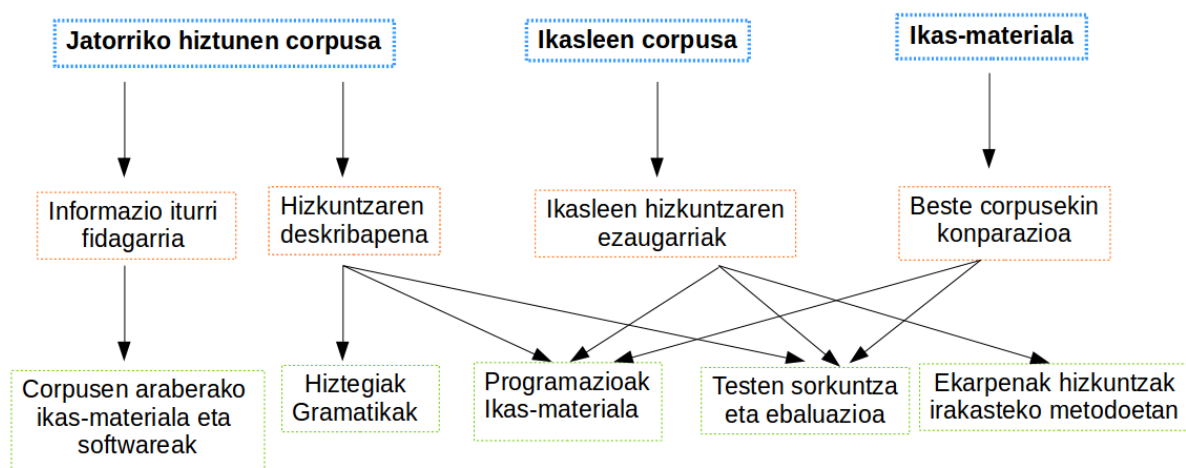
Corpusen erabilerak, beraz, hainbat ekarpen egin ditzake hizkuntzen i(r)a kaskuntzaren alorrean (2. irudia). Eta horretarako hiru corpus-mota bereizten dira:

- *Jatorrizko Hiztunen Corpusak* hizkuntzaren erabilera zuzenaren isla izan ohi dira, eta hizkuntza zehatz eta ongi deskribatzeko aukerak eskaintzen dituzte. Horregatik, hiztegiak edota ikasleentzat egokitutako gramatikak osatzeko oso lagungarriak dira, baita ariketak sortzeko ere.
- *Hizkuntza Ikasleen Corpusak* oso baliagarriak dira ikasleen hizkuntza prozesua ezagutzeko (Granger *et al.*, 2002). Izan ere, ikasleek dituzten beharrean inguruko informazioa ematen dute, eta horrek aukera ematen du ikasleen beharren arabera hiztegiak, gramatikak edota ariketak sortzeko.
- Eskoletan erabiltzen den *Ikas-Materialean osatutako Corpusak* aukera ematen du hizkuntza-ikasleekin erabiltzen den hizkuntza aztertzeko. Ikas-materialetan dagoen hizkuntza eta jatorrizko hiztunena aldera daiteke, ikas-materialek ikasleen erabilera erreala islatzen duten edo aztertzeko, eta horren arabera ikas-material eraginkorrak garatzeko. Mindtek (1996:232) dioten moduan,

“A simple yet important role of corpora in language education is to provide more realistic examples of language usage that reflect the complexities and nuances of natural language... The use of grammatical structures in textbooks for teaching English differs considerably from the use of these structures in L1...”. (Mindt 1996: 232)

10 Gero eta ikastetxe gehiagotan erabiltzen dira, jada, Moodle plataformak (edo Google Classroom), non ikasleek idazten dituzten idazlanak formatu elektronikoan gordetzen diren. Horrek ikasleen testu-bilduma digitalekin lan egitea ahalbidetzen du.

2. irudia: Corpusek hizkuntzen i(ra)kaskuntza-prozesuan egin ditzaketen ekarpenak (McEnery and Gabrielatos, 2005 lanetik egokituta).



Beraz, corpusak lagungarriak izan daitezke ikas-materialak egiaztatzeko, aberasteko, osatzeko edota ikas-materialak hizkuntzaren erabilera errealeran hurbiltzeko, eta ondorioz, erabilgarriak dira hizkuntzen i(ra)kaskuntzarako hainbat ikuspegiatik eta askotariko ekarpenak egin ahal izateko.

Esaterako, eta adibide gisa, Cobben (2007) arabera ikasleak diziplina jakin bati dagozkion kontzeptu eta hitz-gako nagusiak ikasiz gero, diziplina horretan erabiltzen diren hitz esanguratsuenak ikas daitezke [eta baita hitz horiekin lotutako gramatika-egiturak](#).¹¹

The good news is the 2000 list and the AWL together, a combined list of 2570 words, can bring the coverage of an academic text up to approximately 90%. In other words, if you know the first 2000 plus 570 AWL words, then you know about 90% of the words you will meet in any academic text. (Cobb 2007):

Egun, jada, corpusen analisiak baliabide informatikoen laguntzaz egiten dira, eskuz baino errazago eta azkarrago. Hala ere, tresna automatikoen emaitzak ondo ezagutu behar dira beti, eta emaitzak konprobatu (eta atazaren arabera zuzendu) egin behar dira, ez soilik tresna automatikoek akatsak egiten dituztelako, baizik eta hizkuntzak eta erabilerak azalpen sakonagoa izan dezaketelako. Corpusak hizkuntza-teknologiekin aztertuz, askotariko datuak eskuratzen ditugu, tresna bakoitzak eskaintzen dituen aukeren arabera.

Azkenik, hizkuntzak i(ra)kasteko corpusak eskolan erabiltzeari dagokionez, bi modu daude (Leech, 1997):

1. Irakaslea da corpora aztertzen duena eta corpusean analisiak egiteko tresnak erabiltzen dituen (alegia, corpusen analisisa egiteko gaitasuna duena). Irakasleak aztertzen ditu corpusetako adibideak eta prestatzen ditu ariketak, corpusetako informazioa kontuan hartuta. Ikasleek corpusetatik eratorritako ariketa edota material horiekin lan egiten dute gero (Bernardini, 2004), horrela komunikazio errealean erabilitako testuak erabiltzeko motibazioa da ikasleek modu esanguratsuagoan ikastea.
2. Ikasleek ere corpusekin lan egiteko aukera daukate eta ezagutzen dituzte corpusak aztertzeko baliabide informatikoak. Testu-analisiak egin ditzakete landu nahi dituzten

11 [Lan honen helburuetatik at dago proposamen metodologiko bakoitzaren irismena ebaluatzea.](#)

egituren inguruko informazioa bilatzeko edota beraien dokumentuak beste ikasle batzuen lanekin alderatzeko. Horrela, alde batetik, ikasleak protagonismoa hartzen du eta berak nahi duena ikasteko aukera du bere ikaskuntza planifikatuz eta, bestetik, irakasleak laguntzailearen eginkizunak hartzen ditu: corpusak nola prestatu diren gainbegiratzen du edota corpusak nahiz praktika egokiak erakusten ditu.

Hala, eta ikasgaiaren helburuen arabera, irakaskuntza prozesuan irakasleek soilik har dezakete parte edota, zergatik ez, ikasleek ere har dezakete parte bere ikasketa-prozesuan, irakaslearen eta baliabide informatikoen laguntzarekin. Edota modu konbinatuan ere egin daiteke hau, helburuak, materialaren diseinua eta ikasketa-prozesuaren kudeaketa irakasle eta ikasleen artean adostuz.

Bakoitzaren helburu edota beharren arabera, beraz, askotariko azterketak eta esperimentuak egin daitezke corpus elektroniko eta baliabide informatikoen laguntzaz. Hezkuntza-sisteman hizkuntza-teknologiak eskaintzen dizkiguten aukerak ezagutaraztea eta CLARIN-K zentroaren laguntzaz edo aholkularitza baliatuz erabiltzen ausartzea da falta zaiguna.

4. NORBERE TESTUAK AZTERTZEKO TRESNAK

Asko dira, egun, testu-analisirako eskuragarri dauden tresnak. Guk gehien ezagutzen ditugunak dira: i) CLARIN-K azpiegituran sorturikoak: ANALHITZA¹², ContaWords¹³ eta DContado¹⁴, eta ii) bestelakoak: Voyant Tools¹⁵, AntConc¹⁶ eta LanksBox¹⁷. Nahiz eta tresna horiekin (eta beste batzuekin) antzeko gauzak egin daitezkeen, batzuetan hizkuntza edo ataza zehatz baterako komenigarriagoa da tresna bat bestea baino. Edonola ere, aipaturiko tresna horiekin egin daitezkeen guztia egin daiteke web-zerbitzu bidez, honako bi tresna hauekin: ANALHITZA eta Voyant Tools. Horiek horrela, lan honetan, bi tresna horiek soilik aurkeztuko ditugu.

Testu-analisirako tresna edota software asko hizkuntzarekiko independenteak izan ohi dira, hau da, edozein hizkuntzatan idatzitako testuak aztertzeiko gai, karaktereetan oinarritzen direlako (esaterako, Voyant Tools). Baina euskara hizkuntza eranskaria izanik, tresna horiek ezin dute gure morfologia aberatsa aztertu, eta ematen dituzten emaitzak, ondorioz, oso azalekoak dira. Esaterako, ez dituzte hitzak lematizatzen (hitz-formen analisira mugatzen dira), eta lematizazioa urrats garrantzitsua da euskaraz; izan ere, hitz bat modu askotara deklinatuta egoteak ez du esan nahi hitz ezberdina denik.

Hutsune horren aurrean, UPV/EHUko IXA ikerketa-taldean, eta CLARIN-K ikerketarako azpiegituraren barruan, ANALHITZA (Otegi *et al.*, 2017) tresna garatu da testu-analisirako (3. irudia).

12 <http://ixa2.si.ehu.es/clarink/analhitza.php?lang=eu>

13 <http://contawords.iula.upf.edu/>

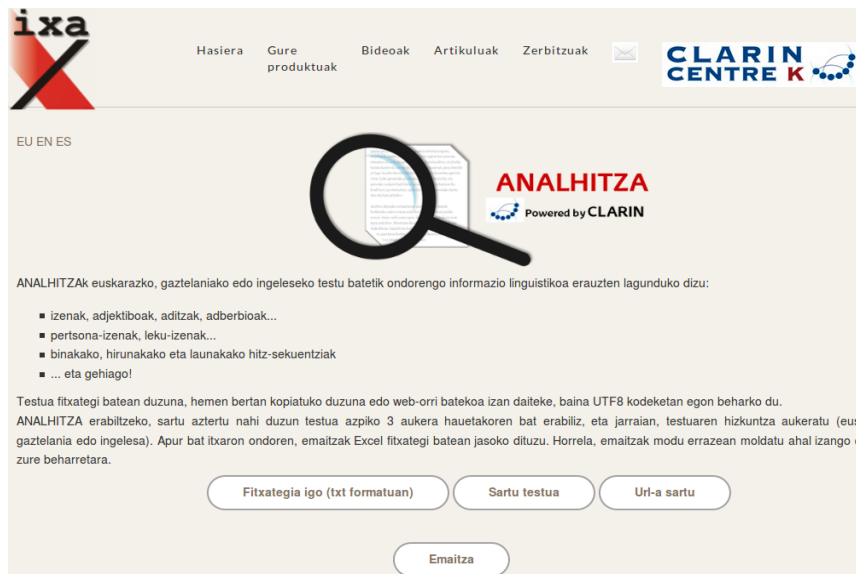
14 <http://sli.uvigo.gal/dcontado/>

15 <https://voyant-tools.org/>

16 <http://www.laurenceanthony.net/software.html>

17 <http://corpora.lancs.ac.uk/lanksbox/>

3. irudia: ANALHITZA tresnaren interfazea.



3. irudia: ANALHITZA tresnaren interfazea.

ANALHITZAk ~~testuak~~ analizatzaile morfosintaktiko batekin analizatzen ditu testuak, hitzak lematizatuz (hau da, lema eta deklinabideak bereiziz). Horri esker, informazio xehatua itzultzen du eta bilaketa aurreratuak egiteko aukera daukagu. Adibidez, testu batean *aita*, *aitak*, *aitari* eta *aitarekin* hitzak baldin baditugu, ANALHITZA gai da *aita* lema eta atzizkiak *-a* (absolutiboa), *-k* (ergatiboa), *-ri* (datiboa) eta *-rekin* (soziatiboa) bereizteko eta konputatzeko bakoitza bere aldetik. Beste hizkuntza batzuetan garatu diren tresnek (Voyant Tools, esaterako), aldiz, hitz desberdin gisa aztertuko lituzkete hitz horiek guztiak, ez baitira gai hitz horietan lema berdina identifikatzeko.

Horrez gain, ANALHITZAk kalkulu-orri batean itzultzen dizkigu aztergai ditugun testuen ezaugarri linguistikoak: testuan dauden hitzak eta hizkiak, esaldi kopurua, esaldiko hitz-kopuruaren batezbestekoa, hitz guztien lema eta hitzak kategoriak sailkatuta (izen, aditz, adjektibo, adberbio, juntagailu, izen bereziak, besteak beste...) ~~sailkatuta, besteak beste~~. Katetoria bakoitzari dagozkion hitzak dokumentu barruko kalkulu-orrietan zerrendatzen dira. Eta hitz mailako informazioaz gain, informazio morfologikoa ere ematen du ANALHITZAk eta N-gramen¹⁸ atalean 2-gramak, 3-gramak nahiz 4-gramak ere ~~ematen ditu ANALHITZAk~~. N-gramen atalak, gainera, garrantzitsuak dira egitura jakinen inguruko adibideak bilatzeko. Informazio horri esker, hain zuzen, 2. atalean ikusi dugun EPD corpusean egin daitezkeen bilaketa modukoak egin daitezke hitz bat baino gehiagoko egiturak aztertzeke. Hala, 2-grama, 3-grama eta 4-gramen orrietan, edozein zutabetan (hitzen zutabeetan, kategorien zutabeetan, edota konbinatuz) aplika daitezke iragazkiak, hitz/kategoria batek aurretik, erdian edota atzetik dituen hitzak zeintzuk diren ikusteko¹⁹ (4. irudia)²⁰.

18 N-grama esaten diogu hitz edo lema sekuentzia bateko eta n hitzez edo lemaz osatutako azpi-sekuentziari.

19 Kontuan izan aditz-laguntzaileen kasuan, lema dela itzultzen duen emailta eta ez forma.

20 Nahi izanez gero, bigarren zutabe horretan beste iragazki bat aplika daiteke aitaren atzetik dauden adjektiboak edota aditzak, esaterako, zeintzuk diren bistaratzeko.

4. irudia: ANALHITZAko kalkulu-orrian aplikaturiko iragazkiaren adibide bat: 2-gramen orrian *aita* (N) hautatu eta atzetik dituen hitzen zerrenda, maiztasunekin eta kategoriekin (V = aditzak, N = izenak, G = adjektiboak, R = izen bereziak, A = adberbioak, O = puntuazio-ikurra, C = juntagailuak/lokailuak, D = determinatzaileak).

A	B	C	D	E	A	B	C	D	E
42	ukan	V	.	O	naiztas	hitza	kategor	hitza	kategor
35	Ordenatu gorantz			O	7	aita	N	etxeokandere	N
31	Ordenatu beherantz			C	2	aita	N	despistatu	V
25	Lehen 10ak			O	1	aita	N	lagundu	V
17	Hutsik			V	1	aita	N	ezagutu	V
17	Hutsik			V	1	aita	N	marisukalde	G
14	Hutsik ez			V	1	aita	N	mantal	N
14	Iragazki estandarra...			V	1	aita	N	eta	C
12				V	1	aita	N	etxe	N
11				O	1	aita	N	asmatu	V
10				C	1	aita	N	espazio	N
10				V	1	aita	N	afari	N
10				N	1	aita	N	despistatu	G
10				V	1	aita	N	serio	G
10				R	1	aita	N	egin	V
9				G	1	aita	N	arotz	N
9				N	1	aita	N	nekatu	V
9				N	1	aita	N	esan	V
8				O	1	aita	N	emegilirik	N
8				N	1	aita	N	sukaldari	N
8				C	1	aita	N	sukalde	N
7				V	1	aita	N	lan	N
7				N	1	aita	N	ematz	N
7				O					
7				O					
6	ukan	V	ni	O					

4. irudia: ANALHITZAko kalkulu-orrian aplikaturiko iragazkiaren adibide bat: 2-gramen orrian *aita* (N) hautatu eta atzetik dituen hitzen zerrenda, maiztasunekin eta kategoriekin (V = aditzak, N = izenak, G = adjektiboak, R = izen bereziak, A = adberbioak, O = puntuazio-ikurra, C = juntagailuak/lokailuak, D = determinatzaileak).

ANALHITZAren landu nahi diren testuak aztertzea bidaltzeko, hiru aukera daude:

- .txt formatuan dagoen fitxategi bat igotzea;
- testua webguneko testu-kaxan sartzea edota kopiazea;
- webgune bateko URL helbidea jartzea, webgune horretako testuak aztertzeko.

ANALHITZA ez da euskarara mugatzen; aitzitik, gaztelaniaz eta ingelesez idatzitako testuak ere aztertzeko gai da, horrek testu eleaniztunen konparazioak egiteko aukera ere ematen duelarik, besteak beste.

Beraz, ANALHITZAk norberari interesatzen zaizkion corpusak aztertu eta askotariko ondorioak ateratzeko aukera eskaintzen du. Horregatik, uste dugu baliabide egokia eta polita dela hizkuntza i(r)a)ksten edota lantzen den eskoletan²¹ erabiltzeko eta baita Humanitate Digitalak ikerketetan erabiltzeko ere. Gainera, oinarritzko hainbat gaitasun garatzeko aukera ere ematen die ikasleei:

- hezkuntza gaitasun orokor batzuk: informazioa interpretatzea, sortzea eta ebaluatzea; ikasteko eta lan egiteko estrategia berriak eskuratzea eta taldean lan egiten ikastea;
- oinarritzko gaitasun batzuk: baliabide berriei ikasten ikastea;
- informazioa tratatzeko teknologia digitala erabiltzea; autoebaluatzen eta kritikoak izaten ikastea;
- gaitasun linguistiko batzuk: testu baten edo gehiagoren kalitate linguistikoa aztertzen jakitea; datu linguistikoak konparatzen, ondorioztatzen edota aurkezten jakitea.

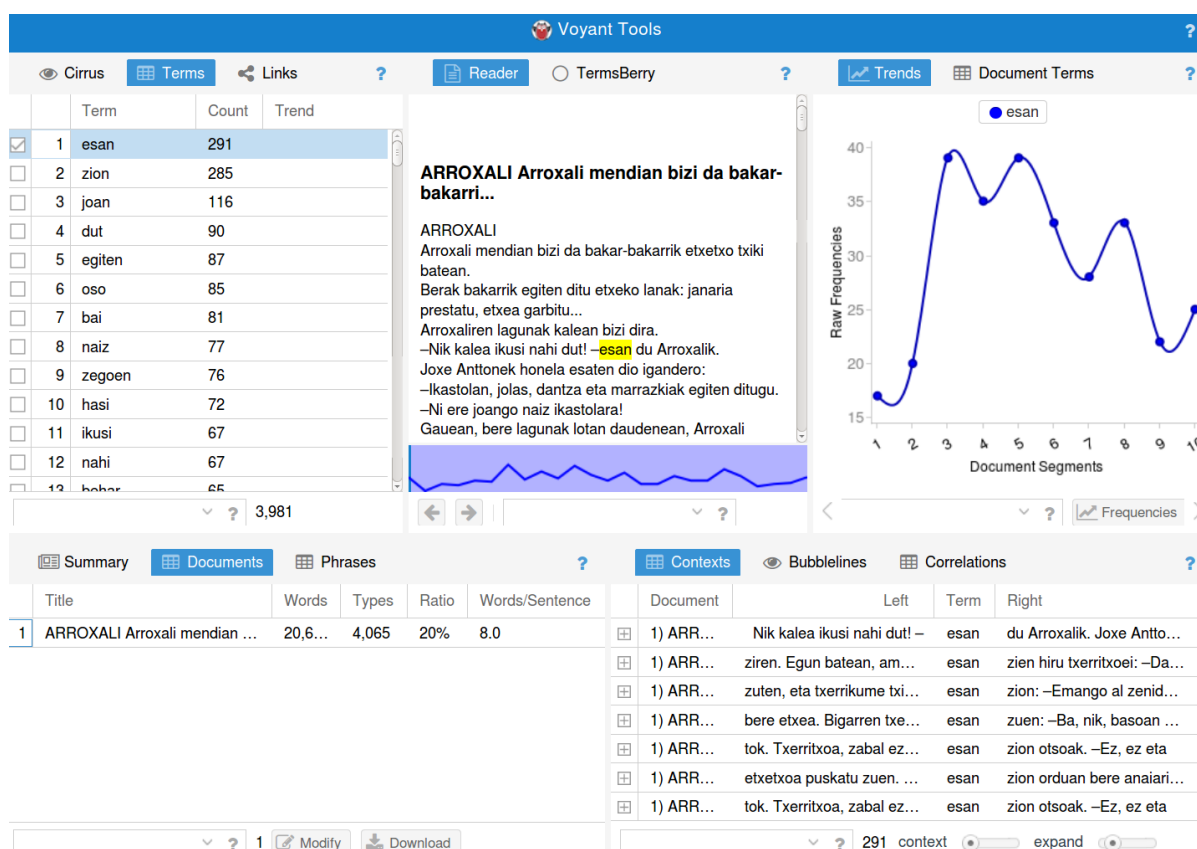
ANALHITZA garapenean dagoen tresna da eta baditu hainbat muga edo hutsune oraindik, batez ere, datuen bistaratzeari dagokionez. Izan ere, kalkulu-orri batean itzultzen ditu emaitza guztiak, eta datuen azterketarako formatu hori oso baliagarria den arren, zenbait erabiltzaileentzat ez dateke erakargarria.

²¹ LHko azken zikloko, DBHko, Batxilergoko eta batez ere unibertsitateko ikasle nahiz irakasleentzat ikusten dugu egokia.

Voyant Tools tresnak, ordea, datuak bistartzeko hamaika programa ditu eta horregatik, batez ere, testu-analisirako beste tresna baliagarri bat da Voyant Tools²². Hori ere Humanitate Digitalen alorrean garatu da corpusen datuak (linguistikoak eta estatistikoak) eskuratzeko.

Kasu honetan, testua aztertu eta emaitzak atal desberdinetan bistartzen dira, bakoitzean informazio mota bat erakutsiz. Esaterako, 5. irudian ikusten dugun moduan (ezkerretik eskuinera eta goitik behera irakurriz), *Arroxali* ipuinaren testua adibidetzat hartuz, Voyant Tools tresnak testu bateko hitzen zerrenda itzultzen du maiztasunaren arabera antolatuta (1); testua osotasunean irakur daiteke, testuingurua galdu gabe (2); hitz bat aukeratu eta dokumentu bakoitzeko hitzaren maiztasunaren grafikoa erakusten du (3); testuaren informazio orokorra ere ematen du (4) eta azkenik, aukeratutako hitzen KWIC (*Key Word In Context*) zerrendak ateratzen ditu (5). Halaber, hitzen lainoak (*Bubblelines*) eta korrelazioak (6. irudia) ere sortzen ditu.

5. irudia: Voyant Tools teknologiak testu baten inguruan ematen duen informazioaren adibidea.



5. irudia: Voyant Tools teknologiak testu baten inguruan ematen duen informazioaren adibidea.

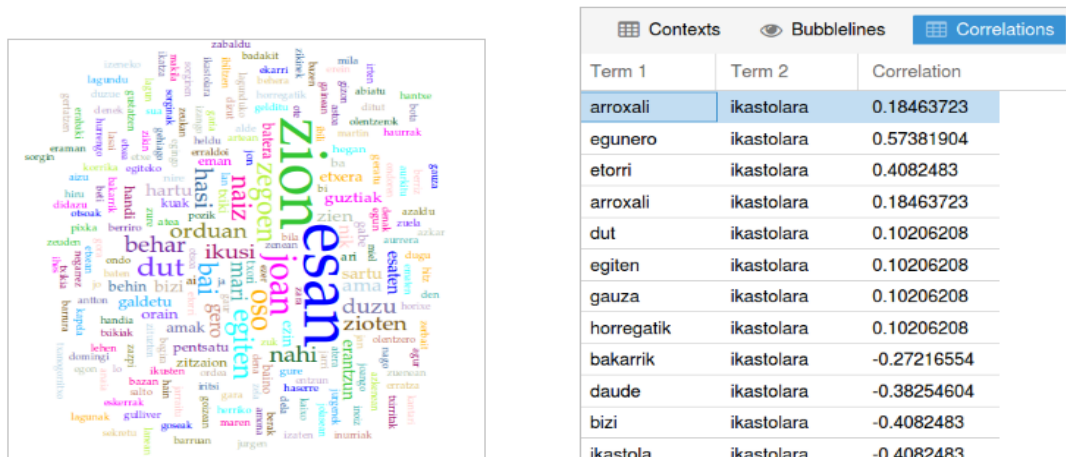
Voyant Toolsek, besteak beste, KWIC (*Key Word In Context*) zerrendak itzultzen ditu, oso lagungarriak direnak klik batekin hitz baten testuingurua zein den ezagutzeko, hau da, hitzaren aurretik eta atzetik zer dagoen jakiteko. Beheko taulak (1. taula), esaterako, gure corpusaren aukeratu dugun lagin batean (ikus 2. taulako 3 urtekoen lehen hiruhileko ipuinak) *behar* hitzaren testuingurua erakusten digu, aztergai dugun hitzaren eskuinean edo ezkerrean dauden hitz aken modu errazean eta goitik behera irakur baititzakegu. Taula horretan ikusten dugu, besteak beste, hiruhileko horretan *behar* hitzarekin osatutako egiturak zein diren eta non dauden (4 ipuinetatik bitan agertzen dela egitura hori eta batez ere 0 Doc delako ipuinean, “Ametsak airean”, hain zuzen ere).

22 <https://voyant-tools.org/>

Doc.	Left	Term	Right
0	Olentzerori haurren eskutitzak eraman	behar	dizkiogu eta!
0	–Olentzerorengana joan	behar	da, eta joan egingo naiz
0	telebistan esan dute elurra egin	behar	duela. Maisua kezkatuta zegoen.
0	dakizu hartzok neguan lo egin	behar	dugula?
0	egia da, lo asko egin	behar	duzue. Barkatu esnatzeagatik.
2	–Bai, bai, hartu	behar	duzun guztia.

1. taula: Ipuinez osatutako gure corpus zatian (3 urtekoen 1. lauhilekoan) “behar” hitzaren erabilerari dagozkion adibide batzuk, KWIC formatuan.

6. irudia: Aztertzen diren testuen hitz-lainoak eta hitzen arteko korrelazioak.



Beraz, Voyant Tools eta ANALHITZA testu-analisiak egiteko bi tresna erabilgarri dira. Voyant Toolsek ANALHITZAK baino datuen bistaratze erakargarriagoa eskaintzen du: datuak formatu edo errepresentazio desberdinetan erakusten ditu, modu txukun eta koloretsuan. ANALHITZAri hori falta zaio, datuak kalkulu orri batean itzultzen dituelako. Voyant Toolsek, baina, testuak hitzen formaren arabera analizatzen ditu, kontaketa ezberdinak eginez eta datuak erakusteko hainbat programa erabiliz, baina informazio morfologikoa edota morfosintaktikoa kontuan hartu gabe. Beraz, atazaren arabera tresna bat edo beste izango da egokiena.

Euskarazko testuen analisirako, baina, ANALHITZAren emaitzak osatuagoak eta xehatuagoak dira, eta, beraz, guk hurrengo atalean aurkeztuko dugun lanerako ANALHITZA bakarrik erabili dugu.

5. CORPUSEN ANALISIA, ANALHITZAREKIN

ANALHITZAren egin daitezkeen analisiak eta ekarpenak zeintzuk diren erakusteko, tresna hau baliatuz egin ditugun bi analisitik atera ditugun ondorioak ekarri ditugu hona:

- Ikastolen Elkarteko HHko ipuinen corpora aztertu dugu, corpus horrek barne hartzen duen informazio linguistikoa (batez ere lexikoa, maiztasuna eta hitzen sekuentziak) zein den jakiteko (5.1. azpiatala).
- Unibertsitateko ikasleen testuak analizatu ditugu lokailuen erabilera aztertzeko (5.2. azpiatala).

Esperimentu edo ariketa bakoitzean lortu ditugun emaitzak eta horrelako tresna batek testuinguru bakoitzean egin ditzakeen ekarpenak aurkeztuko ditugu hurrengo azpiataletan.

5.1 Ipuinen corpusa

Ikastolen Elkarteak²³ egindako materialetan eta hezkuntza programan garrantzia handia du ipuinak. Haur Hezkuntzan, zehazki, 3, 4 eta 5 urteko haurrei zuzenduriko 32 ipuin baliatzen dituzte (2. taula):

3 URTE	4 URTE	5 URTE
1. hiruhilekoa 1. Arroxali 2. Hiru txerrikumeak 3. Txanogorritxo 4. Ametsak airean	1. hiruhilekoa 1. Koxmeren gezurrak 2. Inurria eta txirrita 3. Zazpi antxumeak 4. Mari Domingi eta Olentzero	1. hiruhilekoa 1. Gulliver Liliputen 2. Artzain gezurtia 3. Olentzeroren oparia
2. hiruhilekoa 5. Izei txikia 6. Ehunzango kaskarina 7. Bi indiar txiki inauterietan 8. Txantxangorri mokofina	2. hiruhilekoa 5. Mari Kanela 6. Bi anaiak eta sorginak 7. Lehoia eta kilkerra 8. Martin txiki eta basajaunak	2. hiruhilekoa 4. Makilakixki 5. Ahatetxo itsusia 6. Ziripot eta Mikel Otxin 7. Erraldoi berekoia
3. hiruhilekoa 9. Bremengo musikariak 10. Azenario erraldoia 11. Ttamtoren opor sekretuak	3. hiruhilekoa 9. Hansel eta Gretel 10. Sasi guztien gainetik 11. Jon zikinen anaia	3. hiruhilekoa 8. Babarrun ale magikoak 9. Galtxagorri langileak 10. Suaren sugarretan

2. taula: Corputa osatzen duten ipuinen zerrenda, urte eta hiruhileko bakoitzeko.

Eta ipuin bakoitzeko 3 dokumentu dituzte: ipuinen bertsio luzeak, laburrak eta ipuin bakoitzari lotutako 10 bat abesti (7. irudia)²⁴.

7. irudia: Ipuin baten bertsio luzea (ezkerrean), laburra (erdian eta letra larriz) eta abestiak (eskuinean), adibide gisa.



Basajaunak, ordea, laster konturatu ziren Martin Txiki engainatu egin zituela, eta haren arzetik joan ziren korrika mendian behera.

—E! Martin Txiki! Garia lapurtu digu! Harrapa dezagun! —esan zuten.

Basajaun batek aizkora bota zion, baina Martin Txiki makurtu egin zen eta aizkora zuhaitz batean sartu zen.

Martin Txiki gaitzainondo baten barruan ezkuratu zen. Basajaunek ez zuten aurkitu ahal izan. Eta bilatzeaz aperturik, handik alde egin zutenen, Martin Txiki herrira jaitzi zen korrika-korrika.

Herrian bere abarretatik gari aleak atera zituen, eta heritar guztiak dei egin zion.

—E! Zatozte! Zatozte denok! Garia lortu du! Garia lortu du!

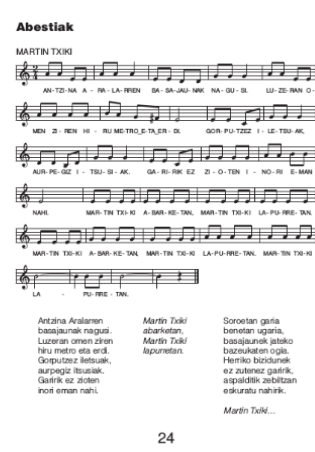
BAINA BASAJAUNAK KONTURATU ZIREN MARTIN TXIKI ENGAINATU ZITUELA.

—MARTIN TXIKI GARI ALEAK LAPURTU DIZKIGU! HARRAPA DEZAGU!

—EZINGO NAUZUE HARRAPATU!

Abestiak

MARTIN TXIKI



Antzina Arelaren basajaunak nagusi. Luzerari onen zion hiru metro eta erdi. Garputzez lekuak, aurpegiz itsuak. Gariak ez zioten inolako emen nahi.

Martin Txiki abarretan. Martin Txiki lapurretan.

Soroetan garia benetan ugari. Basajaunek jaso bazeuketen ogia. Herriko bizidunak ez zuten ez gariak, aspaldik zebiltzan eskuratu nahik.

Martin Txiki...

7. irudia: Ipuin baten bertsio luzea (ezkerrean), laburra (erdian eta letra larriz) eta abestiak (eskuinean), adibide gisa.

Corpusaren tamainari dagokionez, guztira 40.158 hitz ditu corpusak (3. taula):

²³ <https://www.ikastola.eus/>

²⁴ Ipuin bakoitzari dagokion bertsio laburra, luzea eta abesti-sorta aparteko fitxategietan gorde dira kodeketa jakin baten arabera antolatuta (zenbatgarren ipuina, urtea, hiruhilekoa, izenburua eta laburra/luzea/abestia den zehaztuz). Corputa honela sailkatu eta gordetzeak aukera ematen digu interesatu ahal zaizkigun testuekin lan egiteko: labur/luze/abesti guztiekin, urte jakin bateko labur/luze/abesti batzuekin, hiruhileko jakin bateko labur/luze/abestiekin, etab.

Corpusa	Hitz kopurua
Ipuin laburrak	6.061 hitz
Ipuin luzeak	19.917 hitz
Abestiak	14.180 hitz
Guztira	40.158 hitz

3. taula: Ikastolen Elkarteko HHko ipuinen corpusaren banaketa.

Ikastolen Elkarteko ipuinen bilduma ANALHITZAren aztertu da, maila desberdinetan: ipuin labur guztiak, ipuin luze guztiak, abesti guztiak, ipuinen bertsio laburrak urte bakoitzeko, ipuinen bertsio luzeak urte bakoitzeko eta abestiak urteko. Baina lan honen helburua ez da analisi horien emaitza gordinak ematea, ANALHITZAren egin daitezkeen azterketa eta ekarpenen berri ematea baizik.

Artikulu honetan gehienbat ipuin luzeen analisisira mugatuko gara. Ikusi dugu 3hiru, 4lau eta 5bost urte²⁵ko haurrei zuzendutako ipuinen bertsio luzeak²⁵ alderatuz, urte batetik bestera ipuinetan erabiltzen den lexikoan (terminologian) bilakaera bat egon dela. Zehazki, lau urteko ipuinetan dago hitz-kopururik handiena (ipuinak ere luzeagoak direlako), baina hitz desberdinen kopuruari dagokionez, maila honetan daude hitz desberdin gutxien; alegia, hiru eta bost urteko ipuinetan lau urtekoetan baino lema desberdin gehiago dago (%17,26, %17,05 eta %15,31, hurrenez hurren). Beraz, lau urtekoentzako ipuinetan, lema tokenen kopurua igo egiten da baina aniztasuna ez (4. taula).

	Lema tokenak guztira	Lema desberdinen kopurua	Ehuneakoak
3 urte	5.706	985	% 17,26
4 urte	7.680	1.176	% 15,31
5 urte	6.738	1.149	% 17,05

4. taula: Hiru, lau eta bost urteko ipuinetan dagoen hitz-kopurua eta lema desberdin neurriak.

Aipatu datuak biltzeaz gain, ANALHITZAk aukera ematen digu urte bakoitzeko ipuinei dagozkien hiztegiak sortzeko, hitzen zerrendak itzultzen baitizkigu, maiztasunaren arabera antolatuta. Ipuin luzeen corpusean dauden izen ohikoenen zerrenda eskuratu dugu eta, hala, badakigu Haur Hezkuntzako ipuin hauetan gehien erabili diren izenak *etxe* eta *ama* direla (5. taula):

	3 URTEA
137	etxe
131	ama
71	Olentzero
70	behar
68	nahi
57	lagun
56	lan
55	txori

5. taula: Ipuin luzeen corpusean izen ohikoenen eta maiztasunen zerrenda-zatia.

Erraz eskura daiteke, halaber, izen ohikoenen zerrenda urte bakoitzeko (6. taula):

3 URTEA		4 URTEA		5 URTEA	
49	etxe	67	ama	36	etxe
44	ama	52	etxe	31	asto

²⁵ Azterketa honetan, ipuin luzeen analisisira mugatu gara (ipuin laburren corpusa txikiagoa delako eta abestien corpusa bereziagoa delako, kantatzerakoan erabiltzen diren soinu eta onomatopeiez josia).

40	txori	46	maria	30	ahatetxo
27	opor	37	jon	28	olentzero
25	behar	36	antxume	27	ku
24	arte	33	nahi	27	erraldoi
22	olentzero	31	antton	27	herri
22	aita	30	sorgin	25	lan
21	sekretu	30	gari	24	su
20	lagun	29	txirrita	23	haur
20	nahi	26	gezur	22	behar
20	ate	25	lan	21	miel
18	ezin	25	kanela	21	otxin
17	barru	23	behar	20	ama

6. taula: 3, 4 eta 5 urteko ipuinetan izen ohikoenen eta maiztasunen zerrendaren zati bat.

Izen ohikoenak dira Lehenengo bi urtetan ama eta etxe 3 eta 4 urtekoen ipuinetan dira izen ohikoenak (haurrak txikiak izanik, babesa adierazten duten izenak biak). Gutxiago agertzen da 5. urteko ipuinetan, berriz, ama 5. urtekoen ipuinetan gutxiago agertzen da.

Hiztegitxo ala zerrenda hauek haurren ahozko corpusetatik erauziriko zerrendekin alderatzea oso baliagarria izan daiteke euskara bigarren (edo hirugarren) hizkuntza duten haurren irakasleentzat, zein hitz irakats ahal dizkieten jakiteko laguntza gisa, adibidez. San Martinek (2017), esaterako, ANALHITZA erabili du HHko ikasleen hiztegiak deskribatzeko.

Bestalde, eta ANALHITZAk testuaren informazioa kalkulu-orrietan ematen duenez, landu ditugun ipuinetan hiru urtetako lema-zerrendak bildu eta beste urteetako lema-zerrendekin alderatuz, erraz jakin ahal izan dugu zeintzuk diren urte bakoitzeko ipuin bilduman aipatzen diren hitz edo lema berriak. Horrela, irakasleak hitz horiei garrantzia eman diezaiake eta horiek gogorarazteko edota lantzeko ahozko ariketatxoak egin.

Horretarako, 7. taula gure kalkulu orriaren lagin txiki baten moldaketa²⁶ da eta honek erakusten digu, adibidez, 3, 4 eta 5 urteko entzako ipuinetan kon lema horiek zein maiztasuna duten (Maiztas. izeneko zutabea) eta ondoren lema horiek maiztasuna beste urtekoen bildumetan badagoen: (1 zenbakiak adierazten du konparatzen den urteko ipuin-bilduman badagoela eta 0 zenbakiak ez dagoela. B beheko adibidean, 7. taulako datuetara mugatuz, *esan* aditza 3 eta 5 urtetako ipuinetan agertzen da).

Maiztas.	3 URTE	4 URTE	5 URTE	Maiztas.	4 URTE	3 URTE	5 URTE	Maiztas.	5 URTE	3 URTE	4 URTE
86	esan	0	1	123	bat	0	0	120	esan	1	0
30	zer	0	0	46	Maria	0	0	36	etxe	0	0
12	asto	0	0	15	galdetu	0	0	13	ahate	0	0
12	pintto	0	0	15	batzuk	0	0	13	bete	0	0

7. taula: 3, 4 eta 5 urteko en ipuinetan dauden lemen konparazioaren lagin bat, adibide gisa.

Gainera, kategorien arabera zerrendak eskuratzeaz gain, N-gramak hitzen segidan edo inguruan dauden hitzak ere ezagutu ditzakegu ANALHITZAk eskaintzen dituen 2-gramekin, 3-gramekin eta 4-gramekin²⁷. Izena (N) + Aditza (V) egituran, *etxe* izenaren atzetik *joan* aditza agertzen da gehien (17 kasutan), *etxera* *joan* egituran, eta *etxean* *egon*, *etxean* *sartu*, *etxean* *geratu* eta *etxera* *itzuli* ere

²⁶ Espazio kontuengatik ez da taula osoa jarri eta moldaketak egin ditugu formulen emaitzak ikusi eta ulertu ahal izateko.

²⁷ Kalkulu orrian iragazkiak aplikatu eta klik batean eskuratu ahalko ditugun interesatzen zaizkigun hitzen aldamenean datozen bi, hiru edota lau hitzak zeintzuk diren.

erabiltzen dira (8. taula)²⁸, Haur Hezkuntzan umeeek etxearekiko behar duten lotura afektiboan nabarmen agertzen da.

17	etxe	N	joan	V
5	etxe	N	izan	V
5	etxe	N	sartu	V
4	etxe	N	geratu	V
4	etxe	N	itzuli	V

8. taula: 2-graman *etxe* izenaren atzetik dauden aditzak eta maiztasunak.

Hitz baten ondotik ez ezik, hitzaren aurretik zer dagoen ere kontsulta daiteke n-gramen atalean.

Adjektiboak dagokienezan, 2-gramak esker jakin dezakegu, oso erraz, zeintzuk diren izen (N) eta adjektibo (G) segida ohikoenak: *txori txiki*, *ahatetxo itsusi*, *erraldoi berekoi*, *erraldoi handi*, *hitz magiko*, *liburu magiko* (9. taula).

27	txori	N	txiki	G
10	ahatetxo	N	itsusi	G
9	erraldoi	N	berekoi	G
6	hitz	N	magiko	G
6	liburu	N	magiko	G

9. taula: Izen (N) eta adjektibo (G) ohikoenak bilatzeko 2-grama.

Edota jakinik corpus honetan *txiki* eta *handi* direla adjektiborik erabilienak, jakin dezakegu adjektibo horien aurretik ze izen agertzen diren (izen (N) / izen berezi (R) + adjektibo (G) segidak 2-grametan bilatuz): *txori txiki*, *Martin txiki*, *erraldoi handi*, *lagun handi*, *esaterako* (10. taula):

27	txori	N	txiki	G
21	Martin	R	txiki	G
7	erraldoi	N	handi	G
5	lagun	N	handi	G
5	hotz	N	handi	G
5	antxume	N	txiki	G

10. taula: 2-grama *txiki* eta *handi* adjektiboaren aurretik dauden izen eta izen berezi (R) ohikoenak.

Bestalde, *otso* hitza ere ohikoa da HHko ipuin eta abestietan, eta otsoarentzat erabili ohi diren adjektiboak zeintzuk diren jakin dezakegu klik batekin. Ipuin luze eta abestietan, adibidez, *gaiztoa* da otsoaren atzetik agertzen den adjektiborik ohikoena baina *beltz*, *maltzur*, *goseti* eta *handi* ere erabili dira (beti ere, izena + adjektibo egituran) (11. taula):

5	otso	N	gaizto	G
1	otso	N	beltz	G
1	otso	N	handi	G
1	otso	N	maltzur	G
1	otso	N	goseti	G

11. taula: *otso* izenaren atzetik datozen adjektiboak, maiztasunekin, ipuinen bertsio luzeetan eta abestietan.

28 N-grametan, oraindik, lema baino ez dira ikusten, eta corpusean ikusi behar da lema bakoitzari dagokion forma zein den. Baina ANALHITZAk garapen-fasean jarraitzen du eta aurreikusita daukagu beharrezkoa dela atal honetan hitzen formak erakustea, bilatutako egiturak zehazki zeintzuk diren jakin ahal izateko, baina bilaketa lemaren arabera egitea.

Beraz, adibideetan erakutsi ditugun moduko datuak lor ditzakegu ANALHITZAren bidez oso modu errazean, eta irakasle edota ikasle bakoitzari interesatu ahal zaizkion bilaketak egiteko hainbat aukera daude.

Ipuinen corpus honetatik atera daitezkeen ondorioak Haur Hezkuntzako hiztegi nahiz egitura ohikoenen ingurukoak dira, eta baliagarriak izan daitezke hezitzaileentzat (maila horretan erabiltzen diren hitz ohikoenez jabetzeko), materialgileentzat (hiztegi horiek kontuan hartzeko hezkidetzaileak diren material berriak prestatzen direnean, eta konprobatzeko ea *otsoa* bera edo beste pertsonaiaren bat beste era batera ezaugarritzen den ~~ikusteko~~) eta ikertzaileentzat, hizkuntzaren erabileraren inguruko azterketak egin ahal izateko, besteak beste. Zehazki, ANALHITZAk itzultzen duen informazioa baliatuz eta emaitzak aztertuz, HHrako egin daitezkeen ekarpen batzuk hauek izan daitezkeela aurreikusten dugu:

- hiztegiak sortzea (hitz espezifikoak edota egitura jakin batzuenak),
- aztertutako testuetan gehien agertzen diren hitz horiekin jolasak edota kantak asmatzea,
- maila honetara zuzenduak dauden testu desberdinak aztertu eta informazio linguistiko hori guztia ~~hori~~ kontuan hartzea material sorkuntzarako (adibidez, *otsoa* beti ~~beti~~ gaiztoa bezala bakarrik azaltzen bada, beste adjektibo batzuk ere txertatzeko),
- HHko bosgarren mailan hasten dira pixkanaka idazketa eta irakurketarekin, eta gaitasun horietan laguntzeko ere baliagarriak izan daitezke hitzen kategoriak eta maiztasunak kontuan hartzen dituzten hiztegi hauek, baita hizkuntzaren jabekuntzan zailtasunak dituzten haurrekin erabiltzeko ere (adibidez, ezezagunak zaizkien hitzak zerrendan identifikatu eta lantzeko).

Bestalde, erabiltzaileek sortzen edo egokitzen dituzten corpusak erabilgarri jar daitezke Interneten hizkuntzaren prozesamenduko tresnak erabiliz, Haur Hezkuntzan erabiltzen diren ipuinetan informazio linguistikoa bilatzeko, hau da, bilaketa parametro batzuk zehaztuz eta horiek betetzen dituzten adibideak modu errazean lortzeko edota beste ipuin-bilduma batzuetako informazioarekin konparatzeko.²⁹

5.2 Unibertsitateko ikasleen corpusa

ANALHITZA erabili dugu, halaber, unibertsitateko ikasleen testuetan lokailuen eta juntagailuen erabilera aztertzeko. Horretarako, “Berrikuntza eta hizkuntzen irakaskuntza” gaiarekin iritzi-testu bat idazteko eskatu zitzaion Bilboko Irakasleen U.E.ko 4. mailako Berrikuntza Minorreko 50 ikasleri. Idazlan horiekin (beti ere ikasleen anonimotasuna gordez), corpus txiki bat osatu dugu eta bi azterketa egin ditugu.

Batetik, ikasleen corpusean dauden lokailuak eta juntagailuak alderatu ditugu Euskal Gramatika Osoa (Zubiri eta Zubiri, 1995) lanean zerrendatzen diren lokailu eta juntagailuekin. Hala, corpusa ANALHITZAren aztertuz, oso modu errazean jakin ahal izan dugu:

- zeintzuk diren ikasleek corpus honetan gehien eta gutxien erabili dituzten lokailuak eta juntagailuak (12. taulan eskuineko zutabean maiztasunaren arabera antolatuta ikusten direnak);

²⁹ Ipuin bilduma horretan zenbait bilaketa egin daitezke honetako helbide honetan:
<http://ixa2.si.ehu.es/clarink/corpusak/ipuinak/>

- aipatu zerrendarekin alderatuta, zeintzuk diren ikasleek corpusean erabili ez dituzten lokailuak eta juntagailuak eta zeintzuk diren corpusean dauden eta zerrendan ez dauden lokailuak eta juntagailuak (12. taula).

12. taulan bigarren zutabean ikasleek erabili dituzten lokailu eta juntagailuen zerrenda (maiztasunen arabera antolatuta) eta bosgarren zutabean Zubiriren eta Zubiriren (1995) laneko lokailu eta juntagailuen zerrenda. Formulatu bat aplikatuta, jakin dezakegu, zein lokailu dagoen beste zerrendan (1 zenbakiarekin adierazita) eta zeintzuk ez (0 zenbakiarekin adierazita)³⁰.

Ikasleen corpusa		Bai / ez	Zubiri eta Zubiri zerrenda	Bai / ez
Maizt.				
792	eta	1	ere	1
113	baina	1	gainera	1
86	ere	1	bestalde	1
71	edo	1	bertzalde	0
37	izan ere	1	halaber	1
35	hala ere	1	orobat	1
35	beraz	1	berebat	0
32	gainera	1	behintzat	1
18	hala nola	0	bestela	0
7	azken finean	0	behinik behin	0

12. taula: Ikasleen eta testu-liburuko lokailu- eta juntagailu-zerrenda (maiztasunen arabera antolatuta).

Ikasle bakoitzaren datuak hartuta, ikasle bakoitzaren lokailu-zerrenda azter daiteke eta kuartiletan (edota beste sistemaren batean) banatu eskola. Horrela, diskurtso-markatzaile gehien eta ezberdinak erabiltzen dituzten ikasleak multzokatuko genituzke, alde batetik, eta diskurtso-markatzaile gutxien erabiltzen dituzten ikasleak, beste aldetik. Horrela, diskurtso-markatzaile gutxien erabili dituzten ikasleek gehien erabili dituztenen zerrendak hartu eta zein falta zaizkien ikus dezakete. Horrez gain, horiek eskolako ariketa horretan nola erabili dituzten azter dezakete, testuetan bertan begiratuta edozein testu editorerekin bilaketak eginda (edo beste maila batean adierazpen erregularrak erabilia).

Horrez gain, datuak ikustarazteko tresnak ere erabil daitezke adibide horiek testuinguruan aurkitzeko, esaterako, Voyant Tools tresnarekin erdiguneko kuartiletako (edo bestelako irizpideen arabeko taldekatzeak eginez) ikasleen diskurtso-markatzaileak ere azter daitezke (13. taula).

Ikas.	Ezker	Term.	Eskuin
I1	eta pentsamoldeek aldaketak jasaten dituzte	ere	. Arrazoi honengatik, aldaketa horiekin bat
I13	paperezkoak izan daitezke, baita digitalak	ere	. Lehenengo mintegian, IKT-ko material
I13	trataeraz barneratu dudan estrategiak	ere	. Azkenik, irakasgai honetan ikasitakoa
I15	zaiola ezta baztertu behar denik	ere	. Argi dago gizarte globalizatu honetan

13. taula: *ere* lokailua eta bere ondoren puntua duten adibideak lehenengo 20 ikasleenak.

13. taula lortuta ikus daiteke I1ek ez duela *ere* lokailua ondo erabili eta bai, ordea, I13k eta I15ek. Modu horretan ikasleek lokailuen erabilera-arauak errepasatu ondoren, beraien (edo taldekideen) akatsetatik ikas dezakete egindako ariketa bereko adibideekin, modu errazean.

Bestetik, lehenengo 20 ikasle horien diskurtso-markatzaile jakin bat hartu beharrean, bi ikasleren idazlanak aldera daitezke beraien testuetan dauden juntagailu eta lokailuak konparatzeko. Izan ere,

³⁰ Benetako zerrenda luzeagoa da, noski. Hona lagin txiki bat baino ez dugu ekarri, adibide gisa.

ANALHITZA oso lagungarria eta baliagarria da testu desberdinak alderatzeko. Aipatu corpusetik bi idazlan hautatu ditugu, bi ikaslek (I2 eta I5) erabili dituzten lokailu eta juntagailu motak eta maiztasunak ezagutzeko. I2ren eta I5en testuak analizatuz, berehala atera dezakegun ondorio nagusia da *eta* dela bi kasuetan gehien erabili den juntagailua, eta I2k I5ek baino juntagailu kopuru eta mota gehiago erabili dituela. Beraz, analisiaren emaitzen aurrean, irakasleak (edota ikasleek beraiek) modu errazean jakin dezakete: i) bi ikasleek saiatu behar dutela *eta* horrenbeste errepikatu beharrean, beste lokailu batzuk (*horrez gain, baita... ere, halaber*, besteak beste) erabiltzen eta ii) I5ek, oro har, juntagailu eta lokailu gehiago erabiltzen saiatu beharko lukeela. Eta horrelako ondorioak lagungarriak dira irakasle nahiz ikasleentzat, bakoitzaren hutsuneak edota ahulguneak zeintzuk diren antzemateko eta horren arabera ariketak prestatzeko aukera proposatu daitezkeen eskaintzen dutelako. Horrez gain, ikasleek ikas dezakete berdinaren artekoek idatzirik testuetatik eta eskolako gaia ikasteko idatzi direnak testuetatik. Modu horretan, gure ustez, ikasketa planifikatzea beste modu batean errazagoan lor daiteke ikasketa planifikatzea eta ikasketa esanguratsua gora lor daiteke izate.

8. irudia: I2k eta I5ek idazlanean erabili dituzten juntagailu eta lokailuen zerrenda eta maiztasunak.

A	B	C
16	eta	
5	ere	
3	baina	
2	horrez gain	
1	ordea	
1	baita	
1	zein	
1	beraz	
1	ezen	
1	hortaz	
1	are gehiago	
1	gainera	
1	alegia	
1	hala nola	
1	honenbestez	
1	hain zuzen	
1	edo	
1	segun eta	

A	B	
21	eta	
2	izan ere	
1	ez ezik	
1	bestalde	
1	ere	
1	baina	
1	hots	
1	gainera	
1	bere	

8. irudia: I2k eta I5ek idazlanean erabili dituzten juntagailu eta lokailuen zerrenda eta maiztasunak.

Artikulu honetan “Berrikuntza eta hizkuntzen irakaskuntza” gaiarekin unibertsitateko ikasleek idatzitako idazlanetan lokailu eta juntagailuen erabileraren azterketatik atera ditugun ondorio batzuk aurkeztera mugatu garen arren, ANALHITZAK hainbat aukera eskaintzen dizkigu bestelako hainbat datu ere modu erraz eta azkar batean aztertzeko. Esaterako, corpus horretan ikusi ditugun izen ohikoenak hauek dira: *hizkuntza, behar, haur, hezkuntza, eskola, irakasle, berrikuntza, ume, ikasle, ikt, baliabide*. Adjektiboak, berriz, honako hauek: *digital, berri, desberdin, handi, egoki, berritzaile* eta *garrantzitsu*. Guztiak ere gaiari oso lotutako terminoak dira, eta horrek erakusten digu gai horri dagozkion kontzeptu eta hitz garrantzitsuenak zeintzuk diren.

Bestalde, ANALHITZari esker modu errazean detekta daitezke ikasleek idazlanetan aipatu dituzten erreferentzia bibliografiakoak zeintzuk diren, izen berezi eta izen-entitateen orrian begiratuta (14. taula). Zerrenda hau abiapuntu bat izan daiteke, esaterako, testuan aipatu diren autoreak berehalakoan identifikatzeko, eta irakasleak jakin dezake berak eskoletan aipatu dituen autoreak bakarrik erreferentziatu aipatu dituzten ala berri batzuk ere gehitu dituzten. Horrez gain, ikus daiteke

modu sistematikoan erabiltzen dituzten erreferentzia bibliografikoak, hau da, soilik lehen abizena erabiltzen duten edo izen-abizenak erabiltzen dituzten.

Maiztasuna	Entitate izenen zerrenda	Entitate mota
5	Barreras	PER
2	Larrigan	PER
1	Amelia Barqun	PER
1	Amelia Sesma	PER
1	Andy Hargreaves	PER
1	Artola	PER
1	Cummins	PER
1	Dean Fink	PER
1	Etxebarria	PER
1	John Dewey	PER

14. taula. Testuan agertzen diren entitate izenen zerrenda.

ANALHITZAk testu-analisirako eskaintzen dituen aukerak, beraz, askotarikoak dira. Bakoitzak interesatu ahal zaizkion testuak azter ditzake eta dituen interes edota helburuen arabera, hainbat ondorio atera. Ondorio horietan oinarrituta, irakasleak, adibidez, ariketa berriak prestatzeko informazioa aurkituko du eta ikasleak berak ere ikasi ahalko du nahi dituen testuekin “jolastuz”.

ANALHITZAk denbora gutxi duen arren, zenbait ikerketa egiteko baliatu izan da: Otegi *et al.*ek (2017) sistema azaldu dute; San Martinek (2017) Haur Hezkuntzako umeen transkribapenak hartu eta gurasoen jatorriarekin talde ezberdinak sortu eta horien ezaugarriak konparatu ditu; Atutxak (2016) laburpenen corpusaren ezaugarriak aztertzeke erabili du; Alkorta *et al.*ek (2016) sentimenduzko hiztegia ebaluatzeke eta hiztegia erauzteke; Imazek (2017) bere tesiko hitz esanguratsuenak eta horien arteko loturak zein diren argitzeke; Garcia-Arriolak (2017) ipuin sexisten eta hezkidetzailen ezaugarriak aztertzeke. Ikerketa eta gradu amaierako lanez gain, UPV/EHUn HEZikt graduondoko eta Bilboko Irakasleen Unibertsitate Eskolako graduko ikasketetako zenbait talde-lan ere egiteke erabili izan da ANALHITZA.

Horiez gain, gure asmoa da ikerketarako ez den eta unibertsitateaz bestelako testuinguruetan edota erabilera erreal batzuetan ere probatzea, hau da, ikastoletan eta eskoletan, hezkuntza formalari dagokionez, baina **baita** hezkuntza ez-formalean edota informalean ere erabiltzea gustatuko litzaiguke.

6. ONDORIOAK ETA ETORKIZUNeko LANA

Artikulu honetan hizkuntzen i(ra)kaskuntzarako corpusek eta baliabide informatikoez eskaintzen dituzten abantailak edo aukerak aurkeztu ditugu, batez ere UPV/EHUKo IXA taldean garatu den ANALHITZA tresna baliatuz.

Euskara hizkuntza eranskaria izanik, garrantzitsua deritzogu testu-analisiak analizatzaile morfosintaktiko batean oinarrituta egotea^{ri}, eta ez hitzen formen arabera^{koak} izatea^{ri}, beste hizkuntza batzuetan garatu izan diren tresnen kasuan bezala, kasu horietan lortzen diren emaitzak ez direlako horren adierazgarriak.

ANALHITZAk hizkuntza lantzeko egin ditzakeen hainbat ekarpenen adibide batzuk aurkeztu ditugu artikuluan, hots, Haur Hezkuntzan erabiltzen diren ipuinen corpora eta unibertsitateko ikasleek idatzitako testuak. Bi testu-bilduma hauen ezaugarriak guztiz desberdinak dira, baina

ANALHITZAk aukera ematen du, hain zuzen ere, bakoitzak nahi dituen testuak analizatzeko, eta hori da azpimarratu nahi duguna.

Ipuinen corpora aztertuz, jakin ahal izan dugu, besteak beste, hitz erabilienak zeintzuk diren gure haurren lehenengo urteetan kontatzen dizkiegun ipuinetan, eta hitzen segidak aztertuz ikusi dugu, esaterako, zeintzuk diren izen ohikoenen atzetik datozen aditzak edota adjektiboak. Unibertsitateko ikasleen testuak aztertuz, berriz, jakin ahal izan dugu zeintzuk diren gehien erabiltzen dituzten lokailu eta juntagailuak eta horietako zeintzuk dauden, ~~edo ez~~, Euskal Gramatika Osoan.

Honekin guztiarekin, gure ondorio nagusia da ANALHITZA tresna lagungarria dela irakasle, ikasle, materialgile edota ikertzaileontzat, ematen duen informazio linguistikoa baliagarria delako, besteak beste, ikasketa prozesuan ikasleen hizkuntzaren inguruko datuak modu errazean ezagutzeko, eta informazio horretan oinarrituta, materialak prestatzeko edota ikerketa beste ikuspegi batekin osagarritzeko.

Esan behar dugu, baita ere, artikulua honetan euskarazko testuetara mugatu garen arren, ANALHITZA gai dela aztertzeko gaztelaniaz eta ingelesez idatzitako testuak ere eta, beraz, hemen planteatu ditugun moduko azterketak egin daitezkeela aipatu hizkuntzak ikasten edo lantzen ari diren ikasleekin ere.

Etorkizuneko lanei begira jarrita, batetik, ANALHITZA Lehen Hezkuntzan eta Bigarren Hezkuntzan probatu eta erabiltzen irakasteko prest gaude, irakaskuntzako testuinguruan ikasleak bere ikaskuntza-prozesuan izan dezakeen baliagarritasuna aztertzeko edota baliagarri izateko behar diren moldaketak egiteko, IXA-CLARIN-K egitasmoaren barruan.

Hala, ANALHITZAren erabilera eta zabalkundea bideratze aldera, LH6 eta DBH4 mailetako irakasle birekin harremanetan gaude. Tresna aurkeztu diegu eta aurreikusi da ANALHITZA baliabide lagungarria izan daitekeela maila horietan ere hizkuntzaren hainbat aspektu i(ra)kasteko. LH6ko irakasleari egokia iruditu zaio, batez ere, kategoria gramatikalak ~~(izena eta adjektiboa bereziki)~~ lantzeko (izena eta adjektiboa bereziki). DBH-mailan teknologiekin ohituta daudenez, baliabide interesgarria iruditu zaio irakasleari ikasleekin erabiltzeko, besteak beste, kategoria gramatikalak lantzeko, lemak eta hitz deklinatuak identifikatzeko (alegia, lemak eta atzizkiak bereizten ikasteko), ikasleen testuen analisiak alderatzeko, eta n-gramen atala hitz-segida jarraiak lantzeko. Epe laburrean aipatu mailetan esperimendu pare bat egitea aurreikusita daukagu. Gainera, Ikastolen Elkartarekin dugun hitzarmenari esker, ikerketa honek jarraipena izango du material berrien sorkuntzan edota metodologia berrien planifikazioan.

Bestetik, ANALHITZA garatze bidean jarraitzen duen tresna da eta baditu hobetu beharreko hainbat hutsune. Esaterako, analizatzaileak errore batzuk egiten ditu eta, ondorioz, analisi guzti-guztiak ez dira beti zuzenak eta tresna automatiko guztiekin bezalaxe, emaitzak neurri batean konprobatzea komenigarria da.

Gainera, datuen bistaratzean hainbat hobekuntza egin behar direla aurreikusten dugu. Batetik, eta aipatu bezala, garrantzitsua iruditzen zaigu n-gramen ataletan egiten diren bilaketetan hitzen formak bistaratzea, eta ez lemak, orain egiten duen moduan. Eta aztertu nahi da, halaber, Voyant Toolsek eskaintzen dituen tresna batzuk (KWIC eta hitz-lainoak sortzeko aukera, adibidez) implementa daitezkeen.

Bukatzeko, nabarmendu nahi dugu Humanitate Digitaletako proiektuetarako eta hezkuntzarako lagungarriak eta berritzaileak izan daitezkeen tresnak garatzen edota moldatzen jarraituko dugula,

bere garaian Idazlagun³¹ proiektuan egin genuen moduan, eta baliabide horiek eskoletan erabiltzen hasteko metodologia bat, corpusak eta materialak prestatu nahi ditugula, IXA-CLARIN K-zentroko helburuei jarraituz.

Eskerrak

Eskerrak eman nahi dizkiogu CLARIN-K zentroari eta IXA taldeko Esther Miranda lankideari, ANALHITZA hobetzeko eskaini digun laguntza handiarengatik.

Gure eskerrak, baita ere, Ikastolen Elkarteari, ikerketa honen abiapuntu izan den Haur Hezkuntzako ipuinen corpus hori gure eskuetan jartzeagatik.

Eta eskerrak, azkenik, Langile Ikastolako Maite Arostegi eta Peña Florida ikastetxeko Eli Izagirre irakasleei, gure tresna ezagutzeko eta probatzeko erakutsi duten interesa eta eman diguten laguntzarengatik.

ERREFERENTZIAK

Aldabe [Itziar](#)-, Arrieta [Bertol](#)-, Diaz de Ilarraza [A-rantza](#), Maritxalar [Montse](#)-, Niebla [Iani](#)re-, Oronoz [Maite](#)- eta Uria [L-arraitz](#) (2008). Hizkuntzaren Tratamendu Automatikoa Euskararen Irakaskuntzan. *BAT Soziolinguistika aldizkaria*, 2008 (I), 66 zk., 61-69 or.; ISSN: 1130-8435

Aldabe [Itziar](#)-, Maritxalar [M-ontse](#), Perez de Viñaspre [O-latx](#) eta Uria [L-arraitz](#). (2012). Automatic Exercise Generation in an Essay Scoring System. *Proceedings of the 20th International Conference on Computers in Education*, pp. 671-673. ICCE 2012, Singapur, ISBN: 978-981-07-4649-0

Aldabe [Itziar](#)-, Maritxalar [Montse](#)- eta Lopez de Lacalle [Oier](#). (2013). EHU-ALM: Similarity-Feature Based Approach for Student Response Analysis. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: *Proceedings of the Seventh International Workshop on Semantic Evaluation* (SemEval 2013). Pages 580—584, Atlanta, Georgia, June 14-15, 2013 <http://www.aclweb.org/anthology-new/S/S13/S13-2097.pdf>

Aldezabal [Izaskun](#)-, Arriola [Jose-Mari](#)-, Diaz de Ilarraza [A-arantza](#) eta [Sarasola](#) [Kepa](#). (2005). *Hizkuntzalaritza konputazionala*. UEU.

Alkorta [Jon](#)-, Gojenola [Koldo](#)- eta [Iruskiet](#)a [Mikel](#). (2016). Creating and evaluating a polarity – balanced corpus for Basque sentiment analysis. *IWoDA16 Fourth International Workshop on Discourse Analysis*, Santiago de Compostela, September 29 th-30 th. (<https://ixa.si.ehu.es/node/4495>)

Atutxa [Unai](#)- eta [Iruskiet](#)a [Mikel](#). (2017). Abstrakzio laburpen-testuak ebaluatze metodoa, Lehen Hezkuntzan: testu luzera, koherentzia erlazioak, gai nagusia. *Tantak*, 53-80. (<https://ixa.si.ehu.es/node/11342>)

Bernardini [Silvia](#). (2004). Corpora in the classroom: An overview and some reflections on future developments. In J.McH. Sinclair (Ed.), *How to use corpora in language teaching* (pp.15-36). Amsterdam: Benjamins.

Cobb [Tom](#). (2007). Why & how to use frequency lists to learn words. Berreskuratuta hemendik: Lextutor website: <http://www.lexutor.ca/research>.

31 <http://idazlagun.elhuyar.eus/tresna/default.aspx>

- Garcia-Arriola; [Eneritz-Arrate](#). (2017). Hizkuntza Hezkidetzaillea eta Sexista: Haur eta Lehen Hezkuntzako ipuinen azterketa, Hizkuntzaren Prozesamenduko tresnak erabiliz. HEZikt Berezko Tituluko proiektua. UPV/EHU eta UEU.
- Gabrielatos; [Costas](#). (2005). Corpora and Language Teaching: Just a fling or wedding bells? *The Electronic Journal for English as a Second Language* (TESL-EJ), Vol. 8, nº 4.
- Granger; [Sylviane S.](#) ~~eta~~ Meunier [Fanny](#), ~~F.~~ (2010). SLA research and learner corpus research: Friend or foe? Paper presented at the 20th Annual *Conference of the European Second Language Association*. Reggio Emilia, Italy, 1-4 September.
- Granger; [Sylviane S.](#), Hung [Joseph, J.](#) ~~eta~~ Petch-Tyson [Stephanie, S.](#) (Eds.) (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins
- Imaz; [Oier](#). (2017). *Deliberative Democracy and Complex Diversity. From Discourse Ethics to the Theory of Argumentation*. UPV/EHU tesia. Hezkuntza, Filosofia eta Antropologia Fakultatea. Donostia.
- Johns; [Tim](#). (1988). Whence and whither classroom concordancing? In T. Bongaerts, P. de Haan, S. Lobbe; ~~and~~ H. Wekker (Eds.), *Computer applications in language learning* (pp. 9-27). Foris.
- Leech; ~~G~~[Geoffrey](#). (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1-23). New York: Addison Wesley Longman.
- Leech; [Geoffrey](#)~~G~~. (2011). Frequency, corpora and language learning. In: Meunier, F.; De Cock, S.; Gilquin, G.; Paquot, M. (Ed.). *A Taste for Corpora*. In honour of Sylviane Granger. Amsterdam/Philadelphia: Benjamins.
- Maritxalar; ~~M~~[ontse](#), Díaz de Ilarraza; ~~A~~[rantza](#); ~~eta~~ Oronoz; ~~M~~[aite](#). (1997). From Psycholinguistic Modelling of Interlanguage in Second Language Acquisition to a Computational Model. *Proceedings of Computational Natural Language Learning (CoNLL97)*. In conjunction with ACL'97/EACL'97, 50-59. Madrid, Spain. 11-12th July 1997.
- Maritxalar; [Montse](#). (1999). *Mugarri: Bigarren Hizkuntzako ikasleen hizkuntza ezagutza eskuratzeko sistema anitzeko ingurunea*. Informatika Fakultatea, UPV-EHU, Doktorego-tesia, 1999ko azaroaren 30a.
- McEnery; ~~T~~[ony](#); ~~&~~ Hardie; [Andrew A.](#) (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
- McEnery; ~~T~~[ony](#); ~~a~~[et](#) Gabrielatos; [Costas](#). (2005). English corpus linguistics. In B. Aarts & A. McMahon (Eds.), *Handbook of English linguistics*. London: Blackwells.
- Meunier; [Fanny](#). (2011). Corpus linguistics and second/foreign language learning: exploring multiple paths, *RBLA, Belo Horizonte*, v. 11, n. 2, p. 459-477
- Mindt; [Dieter](#). (1996). English Corpus Linguistics and the Foreign Language Teaching Syllabus. In Thomas, J. And M. Short (eds.). *Using Corpora for Language Research*. London: Longman.

- Otegi ~~Arantza~~, Imaz Oier, Díaz de Ilarraza ~~Arantza~~, Iruskieta Mikel ~~eta~~ and Uria ~~Larraitz~~. (2017). ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. *Procesamiento del Lenguaje Natural* 58, pp. 77-84. ISSN: 1135-5948.
- San Martín, Ane. (2017). Haurren ahozko kontakizunen hizkuntza gaitasuna: D ereduko eta euskarazko familiarik gabeko 5 urteko klasean. Gradu Amaierako Lana Haur Hezkuntzako graduan. Bilboko Irakasleen U.E.
- Uria, ~~Larraitz~~, Maritxalar, ~~Montse~~ eta Zabala, ~~Igone~~. (2010). Euskarazko erroreak eta desbideratzeak aztertzeke lan-ingurunea. *HIZPIDE aldizkaria* (ISBN: 1136-5714), 74. zk, p. 16-39, HABE, Donostia.
- Uria, ~~Larraitz~~, Maritxalar, ~~Montse~~ eta Aldabe, ~~Itziar~~. (2011). Basque Learner Corpus Research. *Learner Corpus Reseach 2011*, Louvain-la-Neuve, 15-17 September 2011.
- Uria, ~~Larraitz~~, Maritxalar, ~~Montse~~ eta Zabala, ~~Igone~~. (2011). An Environment for Learner Corpus Research and Error Analysis: the Study of Determiner Errors in Basque. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, 4 (3), 34-51, July-September 2014, edited by Bin Zou (ISSN 2155-7098, eISSN 2155-7101).
- Villegas, ~~Marta~~, Bel, ~~Núria~~, Gonzalo, ~~Carlos~~, Moreno, ~~Amparo~~ ~~eta~~ and Simelio, ~~Núria~~. (2012). Using Language Resources in Humanities research. In *LREC 2012*, pp. 3284-32888
- Zubiri, ~~Ilari~~ eta Zubiri, ~~Entzi~~. (1995). *Euskal Gramatika Osoa*. Didaktiker, S.A. Bilbo.