Lengoaia eta Sistema Informatikoak Saila

eman ta zabal zazu
universidad del país vasco
euskal herriko unibertsitatea

Informatika Fakultatea

Idiomatikotasunaren karakterizazio automatikoa: `izena+aditza` konbinazioak

Antton Gurrutxaga Hernaizek

Informatikan Doktore titulua eskuratzeko aurkeztutako

TESI-TXOSTENA

Donostia, 2014ko ekaina

Lengoaia eta Sistema Informatikoak Saila

eman ta zabal zazu
universidad del país vasco
euskal herriko unibertsitatea

Informatika Fakultatea

Idiomatikotasunaren karakterizazio automatikoa: `izena+aditza` konbinazioak

Antton Gurrutxaga Hernaizek Iñaki Alegriaren eta Xabier Artolaren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua

Donostia, 2014ko ekaina

Tesi-lan hau Elhuyar Fundazioaren *KONBITZ – Hitz anitzeko unitateen eskuratze automatikoa: hiztegi konbinatorioak eratzeko teknikak* eta *KONBITZ2 – Hitz anitzeko unitateen eskuratze automatikoa: idiomatikotasunaren karakterizazioa* ikerketa-proiektuen testuinguruan egin da. Proiektu horiek Eusko Jaurlaritzaren Ekonomia Garapena eta Lehiakortasuna Sailaren SAIOTEK 2011 eta SAIOTEK 2012 ikerkuntzarako programen laguntza jaso dute.

Karmeleri

Eskerrak

Pertsona askok lagundu didate tesi-lan hau egiten, eta guztiak ditut gogoan eskerrak emateko orduan.

- Iñaki Alegria eta Xabier Artola zuzendariak, nigan konfiantza jarri duzuelako, eta honainoko bidean maisu onenen moduan gidatu nauzuelako. Ikerketan barrena sortu zaizkidan bidegurutzeetan, aukera onena identifikatzen lagundu didazue, eta helburura begira jarri beti.
- Ixa taldeko Ainara Estarrona eta Larraitz Uria, eta Elhuyar Fundazioko Nerea Areta eta Ainara Ondarra, ebaluaziorako erreferentzia sailkatua osatzeko egindako lanagatik. Zuen parte-hartzea giltzarria izan da ikerkuntza honetan.
- Elhuyarreko I+Gko Iñaki San Vicente, Perl lengoaiaren inguruko nire zalantzak argitzen eta arazoak askatzen laguntzeagatik, esperimentuen diseinuaz egindako ohar zorrotzengatik, eta malgutasun lexikalerako glosario distribuzionala eratzen erakusteagatik.
- Elhuyarreko I+Gko Maddalen Lopez de la Calle, Lemur Toolkitez dakidan guztia erakusteagatik, eta ikergai honetan aplikatzeko moduaz iradokitako ideiengatik.
- Ixa taldeko Ruben Urizar, ezin konta ahala gai eta alderditan bide-erakusle izateagatik ez ezik, hiztegi-erreferentzia osatzeko EDBLko baliabideak eskura jartzeagatik ere.
- Ixa taldeko Oier Lopez de la Calle, antzekotasun distribuzionaleko esperimentuetan LSA aplikatzen laguntzeagatik, eta Infomap softwarearen erabileraz emandako argibideengatik.
- Ixa taldeko Eneko Agirre eta Elhuyarreko I+Gko Eli Pociello, Euskal WordNet erabiltzeko aukera emateagatik eta aplikatzen orientatzeagatik.

- Ixa taldeko Olatz Arregi, Ikasketa Automatikoan irakasle paregabea izan zaitudalako, eta, zehazki, Weka paketea esperimentuetan aplikatzen gidatu nauzulako.

- Ixa taldeko Aitor Soroa, Latexeko ataka gaiztoetan irtenbidea eskaintzeagatik ez ezik, autonomia handiagoa lortzen ere trebatzeagatik.

- Yosu Yurramendi, estatistikako hainbat erabaki hartzerakoan emandako aholkuengatik. Idiomatikotasuna neurtzeko neurri estatistikoetan eta ebaluazio-metriketan ez nintzen zu gabe inora iritsiko.

- Elhuyarreko I+Gko Igor Leturia, Iker Manterola eta Xabier Saralegi. Ezin hemen aipatu zuengandik ikasitako guztia, batez ere terminoerauzketaz eta corpusgintzaz, eta lagungarri gertatu zaizkidan aholku eta iradokizun guztiak. Bikainak, denak ere. Igor, gainera, I+Gko arduraduna izan da lan hau egin dudan urteetan. Hitz bakarra: *chapeau*!

- Elhuyarreko Hiztegiak+Corpusak ataleko lankideak: Ainara, Amaia, Edurne, Eli, Garbiñe, Klara, Mari eta Sahats. Erraztasun guztiak eman dituzue tesi honek gure jardunean tokia izan dezan.

- Ixa taldeko hainbat kide: Itziar, Nerea, Kike, Gorka, Arantxa eta Montse. Era askotako laguntza jaso dut zuengandik, eta beti prest egon zarete arazoak konpontzeko edo aholku on bat emateko.

- Ixa taldeko Arantza Diaz de Ilarraza, Kepa Sarasola eta Xabier Arregi, ikergai hau Ixa taldearen egitekoen artean kontuan hartzeagatik, eta ikerkuntzan jardutera bultzatzeagatik.

- Elhuyar Fundazioko arduradunak, ikerkuntzaren aldeko apustua egiteagatik. Horri esker egin ahal izan dut lan hau. Batez ere, Josu Aztiria, Hizkuntza & Teknologia unitateko zuzendaria, lan hau etorkizuneko hiztegigintzan erabilgarria izango delako iritzia partekatzeagatik.

- Bereziki dut gogoan Elhuyar Fundazioan urte askoan lankide izan dudan Mariaje Jauregi zena: ikergai hau Elhuyarren jardueraren barruan kokatzeko ikuspegia izan zenuen, eta lan honi ekiteko kondizio ezin hobeak eskaini zenizkidan. Ez daukat ahazteko.

- Lan hau egin dudan bitartean beti ondoan izan zaitudan hori. Karmele, dena erraztu didazu ni honetan aritu ahal izateko, eta egia da: "orrek danak pagatakun ez gazta asko sobrako!"

Eskerrik asko denoi!

Laburtzapenak

AdjS	adjektibo-sintagma
AM	elkartze-neurria (*association measure*)
AP	batez besteko doitasuna (*average precision*)
CCI	zuzen sailkatutako instantziak (*correctly classified instances*)
CPMI	elkarrekiko informazio puntual baldintzazkoa (*conditional pointwise mutual information*)
DS	determinatzaile-sintagma
DSim	antzekotasun distribuzionala (*distributional similarity*)
EDBL	Euskararen Datu Base Lexikala
EH	*Euskal Hiztegia*
ELH	*Elhuyar Hiztegia*
ElhDB	Elhuyarren Datu Base lexikografikoa
HAD	hitzen adiera-desanbiguazioa
HAU	hitz anitzeko unitatea
HAUL	hitz anitzeko unitate lexikala
HB	*Hiztegi Batua*
HP	hizkuntzaren prozesamendua
IE	informazio-erauzketa (*information extraction*)
IR	informazio-berreskuratzea (*information retrieval*)
IS	izen-sintagma
ITA	etiketatzaileen arteko adostasuna (*inter-tagger agreement*)
LF	funtzio lexikala (*lexical function*)
LFlex	malgutasun lexikala (*lexical flexibility*)
LLR	egiantz-arrazoiaren logaritmoa (*log-likelihood ratio*)
LR	egiantz-arrazoia (*likelihood ratio*)
LSA	ezkutuko semantikaren analisia (*latent semantic analysis*)
MI	elkarrekiko informazioa (*mutual information*)

MSFlex	malgutasun morfosintaktikoa (*morfosyntactic flexibility*)
OEH	*Orotariko Euskal Hiztegia*
PMI	elkarrekiko informazio puntuala (*pointwise mutual information*)
PS	postposizio-sintagma
RFR	maiztasun erlatiboen arrazoia (*relative frequency ratio*)
SVD	balio singularretan deskonposatzea (*singular value decomposition)*
SVM	sostengu-bektoreen makina (*support vector machine*)
UF	unitate fraseologikoa
VSM	bektore-espazioaren eredua (*vector space model*)
WN	WordNet
WSM	hitz-espazioaren eredua (*word space model*)
XS	X sintagma

Gaien aurkibidea

Irudien zerrenda

Taulen zerrenda

I. KAPITULUA

Tesi-lanaren aurkezpen orokorra

I.1 Sarrera

Hizkuntza bat ikasteko esperientziatik igaro den orok sentituko zuen maiz zein garrantzitsua den ikastea ama-hizkuntzatzat delako hizkuntza dutenek hitzak nola konbinatzen dituzten, baldin "jatorrizko" hiztunen pareko komunikazio-gaitasuna lortuko badu. Euskarazko *urrats* ingelesez *step* dela ikasita, eta *egin* esateko *do* edo *make* erabil ditzakegula jakinda, gure *urratsak egin* adierazteko, *to do steps* edo *to make steps* konbinazioak sortuko ditugu, modu naturalean, harik eta norbaitek ezetz esan arte, horrela ez dela esaten, ez *do* ez *make*, *step* hitzarekin *to take* erabiltzen dutela ingelesdun "jatorrek". Gaztelaniaz, *dar pasos* erabiltzen da, eta frantsesez, *faire des démarches*. Edo, frantsesez ikasten ari bagara eta *ardo gorri* nola den inork esan ez badigu, jatetxe batean *vin rouge* eskatuko dugu segur aski, baina zerbitzatu orduko jakingo dugu 'ardo beltza' eskatu dugula, *ardo gorri* adierazteko *vin rosé* erabiltzen baita frantsesez. Gaztelaniaz ikasten ari den euskaldun batek ere lasai esango luke *sacar ruido*, gure *zarata atera* hitzez hitz itzulita, baina gehiago ikasi ahala ohartuko da gaztelaniaz *meter ruido* esan ohi dela, aditz antonimoa erabiliz, hain zuzen ere!

Horietan behintzat, *urrats*, *ardo* eta *zarata* hitzek beren "ohiko" esanahia dute, hizkuntza batetik bestera "zuzenean" itzul daitezke; konbinazioaren beste osagaia ez, ordea, eta hizkuntza bakoitzean ohikoa den aditza aukeratzea izango da lanak emango dizkiguna. Baina halako batean norbaitek ingelesez *don't pull my leg* esaten badigu, nekez ulertuko diogu, *pull* eta *leg* hitzen esanahia jakin arren, non ez garen ari une horretan haren hankatik tiraka, edo, ingelesdun onak izanik, esapidearen esanahia zein den aurrez ez

badakigu. Orduan, lagun errukior batek esplikatuko digu 'adarrik ez jotzeko' esan digula. Horrelakoetan, osagaiek ez dute gordetzen beren oinarrizko esanahia, eta konbinazioaren esanahia ikasi ezean, nekez ondoriozta genezake osagaien esanahietatik.

Hitzen konbinatoriaren mundua da hori, hitz anitzeko unitateen (HAU) edo unitate fraseologikoen (UF) arloa, fraseologia. Hor espezie desberdinak bizi dira, hala nola *atentzioa eman*en modukoak, kolokazioak, eta *adarra jo*ren estilokoak, esapide idiomatikoak edo lokuzioak. Fenomeno hau hizkuntza orotan gertatzen da, hizkuntzaren beraren "propietate" unibertsal bat dela uste da (Moon, 1998a). Jakina da hizkuntzaren ezaugarri gakoenetako bat konbinazio-sistema diskretua izatea dela (Pinker, 1994), hau da, multzo mugatua osatzen duten elementu bakunak konbinatuz konbinazio berriak, lehendik inoiz sortu gabeak, era ditzakegula, eta horrexetan datzala hizkuntzaren adierazte-ahalmena (Hauser et al., 2002). Baina ikerketek erakutsi dute hiztunok konbinazio "preferentzial" edo "unitate aurrefabrikatu" batzuk erabiltzen ditugula, unean-unean egindako konbinazio "libreen" gisa berean eratzen ez direnak. Gaur egun, onartua dago hizkuntzaren funtzionamendua ezin dela osagai bakunen konbinazio libreaz (sistemaren gramatikaarauen zein semantikaren arabera) soilik azaldu, hiztunek erabiltzen dituzten hizkuntza-elementu batzuk nolabaiteko unitate "aurrez eratuak" baitira, zenbait osagai bakunez osatutako unitateak izan ere (Fillmore, 1979: 92).

Zenbait autorek enfasi berezia jarri dute hitz anitzeko unitateek hizkuntzan, eta zehazki lexikoan, duten pisuan. Jackendoffek (1995) estimatzen du, telebistako lehiaketa-programa bateko corpusa aztertuta, erabilitako "segida formulaikoen" (*formulaic sequences*) lexikoa hitz bakunen lexikoa adinakoa dela, handiagoa agian. Erman eta Warrenek (2000), ildo beretik, kalkulatu zuten segida formulaikoen proportzioa % 58,6 zela haiek analizatu zuten ingelesezko diskurtsoan. Antzeko baieztapen gehiago ekar genitzake hona, eta intuizioak ere hala iradokitzen digu, nahiz eta ebidentzia enpirikoak ez diren erabat konkluienteak (Schmitt eta Carter, 2004).

Hitz anitzeko unitateek horrenbesteko pisua izaki, zentzuzkoa da uste izatea hizkuntza batean komunikatzen jakiteak, hau da, komunikaziogaitasunak, lotua egon behar duela horiek ezagutzearekin eta erabiltzen jakitearekin. Hizkuntza bat ikasten ari denarentzat, aski zaila da jakitea zein diren, posible liratekeen konbinazioetatik, normalean erabiltzen direnak (Wray, 2000; Warren, 2005). Zenbait ikerketak agerian utzi dute (Howarth, 1998; Wiktorsson, 2003), hizkuntza baten jatorrizko hiztunen eta hizkuntza hori bigarren hizkuntzatzat dutenen jarduerak konparatuta, alde handienetako bat dela bigarrenen fraseologia urriagoa izatea; eta bigarren hizkuntza baten maila desberdineko ikasleen artean ere, maila hobetu ahala handiagoa

dela unitate fraseologikoen erabilera. Zehazki, Howarthek (1998) kolokazioen arloan kokatu du gabezia nagusia.

Horrek guztiak agerian uzten du hitz anitzeko unitateek edo unitate fraseologikoek leku nabaria merezi dutela hizkuntzaren fenomenoa esplikatu nahi duten teorietan, hizkuntzari buruzko informazioa bildu nahi duen edozein hizkuntza-baliabidetan, hala nola hiztegietan eta hizkuntza ikasteko materialetan, eta, azken urteotan argi ikusi denez, hizkuntzaren prozesamendu automatikoan (HP). UFen eskuratze eta prozesatze automatikoa egiteko garrantzitsua da hizkuntzaren teknologiaren hainbat esparrutan: itzulpen automatikoan, IE-IR sistemetan, entitate-erauzketan, terminologia-erauzketan, testu-sorkuntza automatikoan... Fraseologia konputazionalaren arloak interes handia sortu du hizkuntzaren prozesamendu automatikoaren ikerkuntzan ari den komunitate zientifikoan, baita oinarrizko tresnen zein aplikazioen garapenean ari diren ikertzaileen artean ere (Heid, 2008: 341; Krcmár et al., 2013). Nabarmentzekoa da "MWE community" delakoak[1] arlo honetan egiten duena lana, 2003az gero urteroko nazioarteko jardunaldiak antolatuz. Bestetik, UFen erauzketa, ezagutza eta prozesatzearen zailtasuna ere aitortua du komunitate zientifikoak (Sag et al., 2002).

Bada, halako garrantzia izaki, bistan da behar behinenetako bat UFak biltzea dela. Unitate horiek datu-base lexikalean edo hiztegi konbinatorioan zehaztu behar dira, eta, aplikazioaren arabera, behar den informazioa gehitu (esanahia, itzulpena, murrizketa lexikalak, propietate morfosintaktikoak...). Urte askoan, introspekzioa edo eskuzko bilketa izan da hori egiteko modu tradizionala, hiztegigintzan eta baliabide lexikalak eratzeko beste egiteko askotan bezala; azken hamarkadetan, corpusgintzari eta HP arloko teknologia-garapenari esker, prozesu horren automatizazioan urrats handiak egin dira. Automatizazio horrek, batez ere, UFen ezaugarri estatistikoak ustiatu ditu (agerkidetza), kolokazioen erauzketan egin du ekarpen handiena, eta iraultza moduko bat izan zen hainbat hizkuntzatako lexikografian (baita terminologian ere, hitz anitzeko terminoen erauzketan, hain zuzen). UFen beste ezaugarri batzuk kuantifikatzeko ikerlanak nabarmen ugaritu dira azken hamarkadan: ez-konposizionaltasun semantikoa, eta finkapen morfosintaktikokoa zein lexikala. Horrez gain, erauzketa hutsa izan da egiteko nagusia, 2000ko hamarkadaren hasieran esapide idiomatikoen eta kolokazioen arteko bereizketa edo sailkapen automatikoa lantzen hasi zen arte.

Euskararen prozesamendu automatikoaren kasuan, aurrerapausoak bi arlotara mugatu dira, hurrengo atalean zehaztuko dugunez: terminologiaren erauzketa, eta corpusean automatikoki etiketatzea datu-base lexikaletan

[1] http://multiword.sourceforge.net/PHITE.php?sitesig=MWE

adierazita dauden UFak. Are gehiago, euskarazko fraseologia teoriko zein praktikoaren lana lokuzioetara orientatu da bereziki, kolokazioak alde batera utzita (Urizar, 2012). Gainera, ahozko tradizioko lokuzioak eta paremiak izan dira interesgune handiena, batez ere galzorian daudenak biltzeko asmoz.

Baina kontuan hartu behar dugu UFak ez direla multzo itxi bat, eta hizkuntza garatu eta erabili ahala, fraseologia aldatu egiten dela, esapide eta unitate berriak daudela testuetan, eta, horiek eskuratu eta prozesatu gabe, hizkuntza modernoaren prozesamenduak hutsuneak ditu. Esaterako, autore batzuek ohartarazi dute euskarazko komunikabideetan kolokazio berriak erabiltzen direla (Altzibar, 2005), eta orobat esan liteke testu zientifiko, tekniko edo administratiboez, besteak beste (Etxebarria eta Bilbao, 2012).

Horren guztiaren ondorioa da euskararen prozesamendu automatikoak ez duela onurarik atera horrelako unitateen garrantzi handia nabarmendu eta ustiatu duen metodologiatik. Hortaz, fraseologia konputazionaleko arlo garrantzitsu batzuk gutxi garatuta daude; adibidez, kolokazioen erauzketa, eta mota desberdineko UFen eskuratze eta karakterizazio automatikoa.

Tesi-lan honen bidez, ekarpen bat egin nahi izan dugu euskarazko fraseologia konputazionalaren arloan; zehazki, corpusetik `izena+aditza` osaerako unitate fraseologikoak automatikoki erauzteko eta haien idiomatikotasunaren arabera karakterizatzeko teknologiak aplikatzea, garatzea eta esperimentalki ebaluatzea da motibatu gaituen eginbeharra.

I.2 Lanaren kokapena

Hizkuntzaren prozesamenduaren barnean, tesi-lan honen ikertze-arloa fraseologia konputazionala da, eta baliabide lexikalen eskuratze automatikoa da haren testuinguru zehatza. UPV/EHUko Ixa taldearen eta Elhuyar Fundazioaren estrategian, arlo honek hasietatik izan du garrantzia.

Ixa taldeak 25 urte baino gehiago daramatza euskararen prozesamendu automatikorako teknologia garatzen. Hitz anitzeko unitateen prozesamenduak dituen eskakizunez jabetuta, UFen errepresentazioa eta identifikazioa egiteko oinarrizko baliabideak eta tresnak garatu dira. Esaterako, Euskararen Datu Base Lexikala (EDBL) (Aldezabal et al., 2001), zeinetan egitura desberdineko hitz anitzeko unitate lexikalak (HAULak) sartu eta deskribatu baitira (Urizar, 2012). Lokuzioen deskribapenaren funtsa, banako osagaiak adierazteaz gain, *gauzatze-eskema* izeneko ezaugarri-multzoa da; horien artean daude osagai bakoitzaren flexio-murriztapenak, HAULaren aldagarritasuna zehazten dutenak, eta osagaien hurrenkera-aldakuntzak. Batez ere, HAULak testuetan identifikatzea eta etiketatzea da lan horren helburua.

Hori egiteko tresna, berriz, HABIL da (Alegria et al., 2004a). UF jarraituak zein etenak tratatzen ditu, osagaien hurrenkera posible guztiak hartzen ditu kontuan, flexio-murriztapen guztiak betetzen direla egiaztatzen du, eta, azkenik, UFen interpretazio morfosintaktikoak sortzen ditu.

Baliabide eta tresna horiek integratuta daude Ixa taldearen euskararen prozesamendu automatikoaren katean (Aldezabal et al., 2011), MORFEUS tresnan (Aduriz et al., 2000), Eustagger lematizatzailearen aurretik (Ezeiza et al., 1998; Alegria et al., 2002).

Aipatu HAULen artean, entitate-izenak ere egon daitezke. Fernandezek (2012) euskarazko testuetan ageri diren entitate-izenen tratamendurako Eihera tresna garatu du. Horien tratamenduan, bada, lau ataza nagusi bereizi ditu: entitateen identifikazioa, sailkapena (pertsona-, toki- edo erakunde-izena), itzulpena eta desanbiguazioa, entitate-izen bera kategoria batekoa baino gehiagotakoa izan daitekeenean.

Baliabide lexikaletan errepresentatuta dauden hitz anitzeko unitateak identifikatzeaz gain, lan batzuk egin dira unitate berriak testuetan aurkitzeko eta bertatik erauzteko. Terminologiaren arloan izan da batez ere lan hori emankorra, Ixa taldearen eta Elhuyar Fundazioaren lankidetzari esker.

Erauzterm euskarazko terminoen erauzle automatikoa izan zen arlo horretako lehen tresna (Alegria et al., 2004b). Termino bakunakzein hitz anitzekoak erauzten ditu. Tresna hibridoa da, lehen urratsean prozesamendu linguistikoa darabilena testuan termino hautagaiak detektatzeko, eta, bigarrenean, hautagaien informazio estatistikoa darabilena haien terminotasun-rankingak osatzeko. Euskarazko terminologia-erauzketaren ikerketa eta Erauzterm tresna tesi-lan honetan barneratu ez badira ere, haren abiapuntutzat har daitezke. Bestetik, erauzle elebakarren arloan, UZEIk TermiGai tresna aurkeztu du[2].

Bestetik, ELexBI es-eu corpus paraleloetatik termino-bikote elebidunak erauzten dituen tresna da (Alegria et al., 2006), eta Elhuyar Fundazioaren Itzulterm zerbitzuaren oinarria da[3]. TMX formatuko itzulpen-memoriak erabiltzen ditu. Lehenik, hizkuntza bakoitzeko termino hautagaiak erauzten ditu, euskararako Erauzterm erabiliz, eta, gaztelaniarako, UPCko Centre de Tecnologies i Aplicacions del Llenguatge i la Parla (TALP) eta Bartzelonako Unibertsitateko Centre de Llenguatge i Computació erakundeek garatutako Freeling software libreko paketea (Carreras et al., 2004). Ondoren, segmentu bereko hautagaien konbinazioak eratzen ditu, eta hautagai-bikoterik probableenak hautatzen ditu, segmentu-mailako agerkidetzaren eta kogna-

[2] http://www.uzei.com/termigai/
[3] http://itzulterm.elhuyar.org/

tuen informazioan oinarrituta. Egiteko honetarako sortutako beste tresna bat UZEIren LEX2 da[4].

Azkenik, Elhuyar Fundazioak AzerHitz tresna garatu du, es-eu eta en-eu corpus konparagarrietatik termino-bikote elebidunak erauzten dituena (Saralegi et al., 2008). Corpus bakoitzetik termino hautagaiak erauzi ondoren, hautagai-bikoteak osatzen ditu, haien testuinguruen antzekotasun distribuzionalean oinarrituta. Konparatu ahal izateko, testuinguruek hizkuntza berean egon behar dute, eta hiztegiak erabiltzen dira hizkuntza bateko testuinguru-hitzak beste hizkuntzara itzultzeko.

Kolokazioen arloan, adieraziak ditugu aurreko atalean euskararako teknologiak dituen gabeziak. Azkenaldian, nabariak dira hutsune hori betetzeko ahaleginak. Batetik, corpusak kontsultatzeko sistema batzuek eskaintzen dute hitz batekin konbinatzen diren hitzen informazio kuantitatiboa[5]. Bestetik, UZEIk Koloka tresna aurkeztu du[6]. Azkenik, Elhuyar Web-corpusen atariko "Hitz-konbinazioak" atalean[7], tesi-lan honetan garatutako oinarrizko agerkidetza-teknika estatistikoak aplikatu dira 125 milioi hitzeko Elhuyar Web Corpusa prozesatzeko; `izena+aditza` ez ezik, `izena+izenondoa` eta `izena+izena` egiturako konbinazioak ere automatikoki erauzi dira. Erauzketaren emaitzak zenbait neurri estatistikoren arabera ordenatuta kontsulta daitezke, eta corpuseko adibideak bistaratzeko aukera dago.

I.3 Helburuak

Tesi-lan honen helburu nagusia hau da:

Corpusetatik `izena+aditza` osaerako unitate fraseologikoak automatikoki eskuratzeko eta haien idiomatikotasunaren arabera karakterizatzeko teknikak ikertzea, garatzea eta konbinatzea.

Helburu nagusira heltzeko, helburu espezifiko hauek zehaztu ditugu:

1 UFen idiomatikotasunaren definizio operatiboaren zehaztapena. Idiomatikotasuna fenomeno konplexua eta graduala delako oinarritik abiatuta, haren osagai diren propietate neurgarriak zehaztea eta UFen sailkapen-eredua lantzea.

[4] http://www.uzei.com/lex2/

[5] Ikus, esaterako, UPV/EHUko Euskararen Institutuaren *Egungo Testuen Corpusa* (http://www.ehu.es/etc/), Ixa taldearen eta Elhuyar Fundazioaren ZTC-*Zientzia eta Teknologiaren Corpusa* (http://www.ztcorpusa.net), Euskaltzaindiaren *Lexikoaren Behatokia* (http://lexikoarenbehatokia.euskaltzaindia.net), edo *Elhuyar Web-corpusa* (http://webcorpusak.elhuyar.org/cgi-bin/kontsulta.py).

[6] http://www.uzei.com/koloka/

[7] http://webcorpusak.elhuyar.org/cgi-bin/kolokatuak.py

2 UFak automatikoki erauzteko eta karakterizatzeko landu diren teknika linguistikoen eta estatistikoen azterketa konparatiboa, azken urteotako bibliografia zientifikoan oinarritua, gure esperimentuetan aplikatu eta garatuko ditugunak aukeratzeko.

3 UFen erauzketa eta karakterizazio automatikoaren atazak definitzea, bakoitzaren ebaluazio-metodologia zehaztea, eta hartarako behar diren baliabideak (erreferentziak edo *gold standard*ak) eratzea.

4 Ikergaitzat ditugun euskarazko `izena+aditza` osaerako unitateen ezaugarriak zehatz deskribatzea.

5 Idiomatikotasunaren propietate bakoitza neurtzeko lan esperimentala, eta horien emaitzak konbinatzea, ikasketa automatikoko teknika sinpleak erabiliz. Helburua ez da ikasketa automatikoan ikertzea, arlo horretako teknikak ikergai honetara aplikatzea eta haien osagarritasuna aztertzea baizik.

6 Garatu ditugun teknikekin egindako esperimentuak ebaluatzea, eta emaitzak analizatzea. Ondorioak, batez ere, euskarazko baliabide lexikal konputazionalak eratzeari eta hiztegigintzari begira ateratzea. Helburu espezifiko horren barruan, gure interesa da ikergai hauen inguruko ezagutza eskuratzea eta sortzea:

   - Idiomatikotasunaren eta haren propietate bakoitzaren neurketen artean dagoen korrelazioa aztertzea, eta UFen karakterizaziorako teknika eraginkorrenak zein diren ondorioztatzea. UFak erauzteko aplikazioetan, osagaien agerkidetzaren neurketa da teknika estandarra, eta ikertu nahi dugu horren emaitzak hobetu daitezkeen idiomatikotasunaren beste propietateak neurtuz.
   - UFen propietateen ebidentzia enpirikoak zenbateraino datozen bat teoria fraseologikoak UFetarako oro har zein UF kategoria bakoitzerako aurresandakoarekin.
   - Idiomatikotasuna fenomeno konplexua izanik, egiaztatu nahi dugu propietateen kuantifikazioaren emaitzak konbinatuz egindako idiomatikotasunaren baterako karakterizazioa hobea den propietate bereko emaitzak soilik erabiliz egindakoa baino. Horretarako, ikasketa automatikoko teknikak aplikatu ditugu, UFak automatikoki sailkatzeko. Ikaste-prozesuari propietate bakoitzak egiten dion ekarpena zehaztu nahi dugu.

Bestetik, hauek dira gure ikergaia mugatzen duten alderdiak:

- Corpus elebakarrak. UFen itzulpenari eta erauzketari begira corpus elebidunetatik lor daitekeena tesi-lan honetatik kanpo geratzen da.
- `izena+aditza` osaerako konbinazioak hartu ditugu ikergaitzat. Interes handiko konbinazio-mota da, oso baita ugaria, eta oraindaino ez da erauzketa automatikoan ikertu.
- Erauzketa eta karakterizazioa dira hautatu diren atazak, batik bat baliabide lexikalak eratzea eta elikatzea jomuga izanik. Ikerketa honen helmenetik kanpo daude, beraz, UFak testuinguru zehatz batean identifikatzea, anbiguotasunaren azterketa eta testuinguruaren araberako ebazpena, eta itzulpenari begira egindako azterketa.

I.4 Tesi-txostenaren egitura

Tesi-txosten hau kapitulu hauetan dago egituratuta:

1 Tesi-lanaren aurkezpen orokorra

Irakurtzen ari zaren kapitulu honetan, lehenik, ikergaiaren aurkezpen orokorra egin dugu, eta hari ekiteko izan dugun motibazioa azaldu. Gero, ikergaiak euskararen prozesamendu automatikoaren arloan eta hiztegigintzan duen kokapena zehaztu dugu. Jarraian, lanaren helburuak formulatu ditugu. Azkenik, tesi-lan honekin zuzenean nahiz zeharka lotutako argitalpenak zerrendatuko ditugu.

2 UFen idiomatikotasunaren eta haren karakterizazioaren marko teorikoa

Idiomatikotasun terminoari eman zaizkion adierak aurkeztu ondoren, ikerlan honetan aukeratu duguna zehaztuko dugu. Idiomatikotasunaren izaera konplexua azalduko dugu, eta haren definizioan parte hartzen duten propietateak landuko: instituzionalizazioa, ez-konposizionaltasun semantikoa, eta finkapen morfosintaktikoa zein lexikala. Bestetik, karakterizaziorako sailkapen-proposamen bat aurkeztuko dugu. Azkenik, euskarazko fraseologiaren ikuspegi laburra emango dugu, eta `izena+aditza` konbinazioen berezitasunak azalduko.

3 UFen erauzketa eta karakterizazio automatikoa

Kapitulu honen xedea da UFen erauzketa eta karakterizaziorako teknologiaren uneko egoera aurkeztea. Horren helburuak eta urratsak aurkeztu ondoren, hautagaiak erauzteko teknika linguistikoei heldu diegu,

eta, jarraian, bi karakterizazio-ataza bereiziko ditugu, ranking bidezkoa eta sailkapen automatikoaren bidezkoa, eta bakoitzerako ebaluazio-metodologiak aurkeztuko. Kapituluaren zati handienean, karakterizaziorako teknika esperimentaletan barneratuko gara, propietate bakoitzaren neurketarako erabili diren prozedurak azalduz lehenik, eta, bigarren, ikasketa automatikoaren bidez propietate horien emaitza esperimentalak sailkapen-atazan nola konbinatu diren aurkeztuz.

4 Lan esperimentalaren diseinua

Erauzketa eta karakterizaziorako diseinatu dugun estrategia deskribatuko dugu kapitulu honetan: esperimentuak diseinatzeko erabili ditugun irizpideak, bereizi ditugun egitekoak eta karakterizazio-atazak, eta ikergaitzat hartu dugun euskarazko `izena+aditza` konbinazio-motaren zehaztapena.

5 UF hautagaiak erauztea

Erabili dugun corpus-baliabidea eta haren aurreprozesamendu linguistikoa deskribatu ondoren, hartatik UF hautagaiak erauzteko garatu dugun prozedura xehatuko dugu, bi urratsetan: bigrama-sorkuntza eta bigramen forma kanonikoa lortzeko normalizazioa.

6 Ebaluazio-metodologia eta baliabideak

Gai hauek landuko ditugu: ranking bidezko eta sailkapen automatikoaren bidezko atazak ebaluatzeko prozedurak; garatutako baliabideak (hiztegi-erreferentzia eta eskuz sailkatutako erreferentzia); eta ataza bakoitzean erabilitako metrikak.

7 Idiomatikotasuna karakterizatzeko esperimentuak

Lehenik, idiomatikotasunaren osagai diren propietateak bereiz neurtzeko esperimentu bakunak deskribatuko ditugu, eta haien emaitzak aurkeztu eta analizatu. Bigarren, ikasketa automatikoko esperimentuak aurkeztuko ditugu: esperimentu bakunetan lortutako emaitzak nola konbinatu ditugun, lortu ditugun emaitzak eta horien analisia.

8 Ondorioak eta etorkizuneko lanak

1. kapituluan egindako ikertze-galderei erantzuteko, aurreko kapituluan egindako esperimentuen analisitik ateratako ondorioak eta tesilan honen ekarpenak laburbilduko ditugu. Azkenik, ikertze-arlo honetan jorratzekoak liratekeen etorkizuneko bideak azalduko ditugu.

I.5 Argitalpenak

Tesi-lan honekin zuzenean lotutako argitalpenak

- Gurrutxaga, A. eta Alegria, I. (2011). Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, 2-7 or. Portland, Oregon: Association for Computational Linguistics.

- Gurrutxaga, A. eta Alegria, I. (2012). Measuring the compositionality of NV expressions in Basque by means of distributional similarity techniques. *Proceedings of the8th International Conference on Language Resources and Evaluation (LREC 2012)*, 2389-2394 or. Istanbul: ELRA.

- Gurrutxaga, A. eta Alegria, I. (2013). Combining different features of idiomaticity for the automatic classification of noun+ verb expressions in Basque. *Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013) NAACL–HLT 2013*, 116-125 or. Atlanta, Georgia: Association for Computational Linguistics.

Terminologia-erauzketaren arloko argitalpenak

- Alegria, I., Gurrutxaga, A., Lizaso, P., Saralegi, X., Ugartetxea, S., eta Urizar, R. (2004b). Linguistic and statistical approaches to Basque term extraction. *Proccendings of GLAT-2004: The Production Of Specialized Texts*, 235-246. Bartzelona: ENST Bretagne.

- Alegria, I., Gurrutxaga, A., Lizaso, P., Saralegi, X., Ugartetxea, S. eta Urizar, R. (2004c). A XML-based term extraction tool for Basque. *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, 1733-1736. Lisboa.

- Gurrutxaga, A., Saralegi, X., Ugartetxea, S. eta Alegria, I. (2005). Eraузterm: euskarazko terminoak erauzteko tresna erdiautomatikoa. *Mendebalde Kultur Alkartea, IX. Jardunaldiak: Euskera zientifiko-teknikoa*. Bilbo.

- Gurrutxaga, A., Pagoaga, A., Saralegi X., Ugartetxea S. eta Alegria I. (2005) Euskara-gaztelania terminologia elebidunaren erauzle automatikoa. Ugarteburu, I. eta Salaburu Etxeberria, P. (ed.) *Espezialitateko hizkerak eta terminologia II. Euskara estandarra eta espezialitate hizkerak*. Leioa: UPV/EHU.

- Alegria, I., Gurrutxaga, A., Saralegi, X. eta Ugartetxea, S. (2006). Elexbi, a basic tool for bilingual term extraction from Spanish-Basque parallel corpora. *Proceedings of the 12th EURALEX International Congress*, 159-165. Turin.

- Saralegi, X., San Vicente, I. eta Gurrutxaga, A. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008) - Building and using Comparable Corpora workshop*, 27-32 or. Marrakex.

- Gurrutxaga, A., Leturia, I., Pociello, E., Saralegi, X. eta San Vicente, I. (2009). Evaluation of an automatic process for specialized web corpora collection and term extraction for Basque. *ELexicography in the 21st century: new challenges, new applications; proceedings of ELex 2009*, 22-24. Louvain-la-Neuve, Belgika.

- Gurrutxaga, A., Leturia, I., Pociello, E., San Vicente, I., eta Saralegi, X. (2010). Internet, corpusak eta terminologia: Internetetik espezialitate-corpusak erauzteko teknikak eta horien ebaluazioa. Alberdi, X. eta Salaburu , P. (ed.) *Ugarteburu Terminologia Jardunaldiak. Euskararen garapena esparru akademikoetan. Espezialitate hizkerak eta terminologia IV*, 69-82. Leioa: UPV/EHU.

- Gurrutxaga, A., Leturia, I., Saralegi, X. eta San Vicente, I. (2013). Automatic comparable web corpora collection and bilingual terminology extraction for specialized dictionary making. Sharoff, S., Rapp, R., Zweigenbaum, P., eta Fung, P. (ed.) *Building and Using Comparable Corpora*, 51-75. Springer.

HParen arloko argitalpenak

- Areta, N., Gurrutxaga A., Leturia I., Polin, Z., Saiz, R., Alegria, I., Artola, X., Diaz de Ilarraza, A., Ezeiza, N., Sologaistoa, A. Soroa,

A. eta Valverde, A. (2005). Zientzia eta teknologiaren corpusa. Diseinua eta metodologia. Ugarteburu, I. eta Salaburu Etxeberria, P. (ed.) *Espezialitateko hizkerak eta terminologia II. Euskara estandarra eta espezialitate hizkerak*. Leioa: UPV/EHU.

- Areta, N., Gurrutxaga A., Leturia I., Polin, Z., Saiz, R., Alegria, I., Artola, X., Diaz de Ilarraza, A., Ezeiza, N., Sologaistoa, A. Soroa, A. eta Valverde, A. (2006). Structure, Annotation and Tools in the Basque ZT Corpus. In*Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1046-1411. Genoa.

- Areta N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., Díaz de Ilarraza, A., Ezeiza, N. eta Sologaistoa, A. (2007). ZT Corpus: Annotation and tools for Basque corpora. *Corpus Linguistics 2007*. Birmingham.

- Leturia, I., Gurrutxaga, A., Areta, N., Alegria, I. eta Ezeiza, A. (2007). EusBila, a search service designed for the agglutinative nature of Basque. *SIGIR 2007- iNEWS'07 workshop*. Amsterdam.

- Leturia I., Gurrutxaga A., Alegria I., Ezeiza A. (2007). CorpEus, a web as corpus tool designed for the agglutinative nature of Basque. *WAC3 2007 (Web as a Corpus) workshop*. Louvain-la-Neuve.

- Leturia, I., Gurrutxaga, A., Areta, N. eta Pociello, E. (2008). Analysis and performance of morphological query expansion and language-filtering words on Basque web searching. *Proceedings of the 6th International Conference on Language Resources and Evaluations (LREC 2008)*. Marrakex.

- Pociello E., Gurrutxaga A., Agirre E., Aldezabal I. eta Rigau G. (2008). WNTERM: Combining the Basque WordNet and a Terminological Dictionary. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakex.

- Areta, N., Gurrutxaga A. eta Leturia I. (2008). Begiratu bat corpusbaliabideei. *BAT Soziolinguistika aldizkaria*, 62. alea. 71-92.

- Gurrutxaga, A., Leturia, I., Pociello, E., Saralegi, X. eta San Vicente, I. (2010). Exploiting the Internet to build language resources for less resourced languages. *SALTMIL 2010 workshop*. Valetta, Malta.

II. KAPITULUA

UFen idiomatikotasunaren eta haren karakterizazioaren marko teorikoa

Kapitulu honetan, gure lana kokatu dugun marko teorikoa aurkeztuko dugu, eta, horretarako, tesi-lanaren izenburua osatzen duten hiru terminoren zehaztapena egingo: *idiomatikotasun* terminoaren adiera, unitate fraseologikoen karakterizaziorako sailkapen-eredua, eta euskarazko `izena+aditza` osaerako konbinazioen ezaugarriak.

II.1 Idiomatikotasuna teoria fraseologikoan

Ikerketa honetan, UFen idiomatikotasuna automatikoki karakterizatzea hartu dugu xede nagusitzat. Beraz, gure lehen egitekoa da *idiomatikotasun* terminoaren esanahia zehaztea. Izan ere, terminoa erabat baliokide edo sinonimo ez diren kontzeptuak adierazteko erabili da fraseologian eta horrelako unitateez arduratu diren beste zenbait arlotan, eta ezinbestekoa da argi uztea zein adieratan erabili dugun gure lanean.

Idiomatikotasunaren definizio zabalduena ez-konposizionaltasun semantikoarekin identifikatu izan da, edo, propietate diskretua ez baina graduala dela uste dutenen ikuspegitik, konposizionaltasun partzialarekin ere bai. Hala ere, ikertzaile batzuek kontzepzio hedatuagoa proposatu dute, UF izatearen ezaugarri guztiak integratu nahian.

Alderdi horiek guztiak II.1.2 atalean xehatuko ditugu, baina, oraingoz, aurreratu dezakegu gure ikerkuntza honetan bigarren ikuspegi hori hartu dugula, hau da, kontsideratu dugu idiomatikotasuna konbinazio bat UF izatea determinatzen duen propietate konplexua eta graduala dela, ez-konposizio-

naltasunaz edo konposizionaltasun partzialaz gain, beste propietate batzuk ere barnean hartzen dituena, hala nola instituzionalizazioa eta finkapen morfosintaktikoa zein lexikala.

Horren azalpen zehatzari ekin aurretik, gaia testuinguru zabalean kokatuko dugu eta, horretarako, UFen ezaugarriez jardungo dugu hurrengo atalean.

II.1.1 Unitate fraseologikoen ezaugarriak

Eskola eta joera guztiek onartuko luketen baieztapena da fraseologiak hitz-konbinazioez diharduela, baina ez edozein konbinazioz, ezaugarri jakin batzuk dituzten konbinazioez edo "hitz anitzeko unitateez" baizik. Esan genezake, orduan, fraseologia definitzea, hein handi batean, haren aztergai diren unitate horien ezaugarriak zehaztean datzala (Granger eta Paquot, 2008: 27).

Lehen koska unitatearen izendapena bera dugu, termino-ugaritasun handia baitugu[1]. Segur aski, horrelako ugaritasunaren arrazoia ez da beti izendapenaren estandarizaziorik eza izango, besterik gabe; kontzeptuari berari buruzko ikuspegi desberdinak eta adostasunik eza ere tartean egon daitezke. Corpas Pastorrek (1996) hiru familiatan antolatu ditu direlako terminoak:

- *hitz anitzeko unitate* erakoak (edo *multiword expression*)
- *esapide finko* erakoak (*fixed expression*)
- *unitate fraseologiko* erakoak (*phraseological unit*)

Urizarrek azaldu duenez (Urizar, 2012: 54-55), lehen multzoko terminoek UFak hitz batez baino gehiagoz osatuak izatearen ideia jasotzen dute

[1]Ingelesez, honakoak aurki daitezke literaturan: *multiword unit, multiword expression, multiword lexeme, multiword lexical unit, multi-word lexical phenomena, phraseological unit, phraseme, conventional expression, formula, formulaic expression, prefab, composite, fixed expression, set expression, set phrase, word combination, phrasal lexeme*. Gaztelaniaz ere, termino-aniztasuna dago: *expresión pluriverbal, unidad pluriverbal lexicalizada y habitualizada, unidad léxica pluriverbal, expresión fija, unidad fraseológica, fraseologismo, frasema*. Frantsesez, *expression multi-mot, unité plurilexicale, unité phraséologique, phraséologie, phrasème, expression figée*. Euskaraz ere, aurreko lerroetan bertan, *unitate fraseologiko* erabili dugu, ondo gogoan izan arren hori bezain erabiliak direla *hitz anitzeko unitate* edo *hitz anitzeko unitate lexikal* terminoak. *Lokuzio* ere oso erabilia da, baita *esapide* ere. Egia esan, beste batzuetan ez bezala, euskarazko termino-barreiatzea hutsaren hurrengoa da beste hizkuntza batzuenarekin konparatuta. Nolanahi ere, gure lanean horietako zein erabiliko dugun eta zergatik aurreraxeago azalduko dugu.

(hau da, polilexikalitatea); bigarrenekoek, esapideen egonkortasuna iradokitzen dute (finkapenarekin eta zurruntasunarekin ere erlaziona genezakeena); azkenik, hirugarren multzoko terminoez ari dela, Urizarrek dio unitate semantikoa osatzen duten egitura sintaktikoak diren aldetik hartzen dutela izena[2]. Izendapenen hirukoiztasun horrek fraseologiak aztergaitzat dituen unitateen ezaugarrien inguruko lehen ideia batzuk eman dizkigu. Ideia horien osagarri, komeni da sarreran esandakoa gogoraraztea: hiztunok hitz-konbinazio "preferentzial" edo "unitate aurrefabrikatu" batzuk erabiltzen ditugu, unean-unean egindako konbinazio "libreen" gisa berean eratzen ez direnak, ezin baitira sistemaren gramatika-arauen zein semantikaren arabera soilik aurreikusi edo azaldu.

Bestetik, ez dago erabat finkatuta fraseologiaren zerizana zein den, aztertze-eremuaren hedadura norainokoa den (alegia, hitz anitzeko unitate guztiak hartzen dituen), ezta eremu horren sistematizazioak eta kategorizazioak nolakoa behar lukeen ere (Montero Martínez, 2002). Hainbat autorek nabarmendu dutenez (Evert eta Krenn, 2005a; Granger eta Paquot, 2008: 28-29; Seretan, 2011: 11-13), bi ikuspegi edo tradizio nagusi bereizi ohi dira[3]:

Ikuspegi linguistikoa (fraseologikoa ere esan ohi zaiona). Eskola errusiarrak eta horretan oinarritu direnek osatua (Cowie, Howarth, Hausmann, Gläser, Choueka, Corpas, Kjellmer, Nesselhauf, Mel'čuk). Ikuspegi honen ideia gakoak hauek dira:

- Fraseologiaren hedadura linguistikoki definitutako unitate-multzo batera mugatzen du.
- Osagaien arteko erlazio sintaktiko espezifikoa (konbinazio bitarren kasuan, `izena+aditza`, `izena+adjektiboa`, eta abar). Erlazio hori ez da distantziaren bidez zehazten (alderdi bakarra ez da elkarren ondoan edo gertu agertzea), erlazio sintaktikoaren bidez baizik.
- Unitateek irregulartasun semantikoa, sintaktikoa eta distribuzionala dute.
- Fenomenoaren alderdi estatistikoa hartzen du kontuan, batez ere kolokazioen kasuan (konbentzionalak, karakteristikoak edo errekurrenteak direla esan ohi da).

[2]Hala ere, gure iritziz *unitate fraseologiko* terminoak, aurreko biek ez bezala, ez du Urizarrek esleitutako esanahia esplizituki adierazten; ez da esanahi horrekiko gardena edo autodefinitzailea.

[3]Aurrerago ikusiko dugunez, ikuspegi-bikoiztasun hori UF-mota baten kontzepzioan azaleratu da batez ere: kolokazioetan.

- Ereduaren gunean unitate prototipikoak daude (esapide idiomatikoak edo *idioms* direlakoak); periferian, bestelakoak (kolokazioak).
- Fraseologia continuum bat da, konbinazio opako eta finkatuenak mutur batean dituena, eta bestean, gardenen eta malguenak.

Ikuspegi estatistikoa (kontestualista, distribuzionalista edo enpirikoa esan ohi zaiona). Firth eta Sinclairren ideietan oinarritua da. Ezaugarri nagusiak:

- Muineko kontzeptua *agerkidetza* da (*co-occurrence*): hitzak testuinguru berean agertzea.
- Testuingurua zehazteko, distantzia erabiltzen da (*window span*, edo "leiho-zabalera"), ez erlazio sintaktikoa.
- Eredua datuetatik eraikitzen da, enpirikoki, ez aurrez definitutako kategoria batzuetatik.
- Corpusean behatzen diren hitz-konbinazioak ardatz batean koka daitezke, nolabaiteko continuum bat osatuz, Sinclairren *open-choice principle* eta *idiom principle* direlakoen araberako muturren artean (Sinclair, 1991: 110).
- Eredu horretan, *kolokazio* kontzeptua zentrala da, ez periferikoa, kolokazioek edo "maiz gertatzen diren hitz-konbinazioek" (Cruse, 1986: 40) esapide idiomatikoek (*idioms*) baino pisu edo maiztasun handiagoa baitute hizkuntza-erabileran (Moon, 1998b: 79).

Ikuspegi linguistikoak eragin handia izan du fraseologia jakintza-arlo heldu bat izan dadin, oinarri teoriko sendoak ezarri ditu, eta nagusi izan da tradizio fraseologikoan eta Europa kontinentaleko hiztegigintzan. Ikuspegi estatistikoak fenomeno fraseologikoak agerkidetzarekin lotzen ditu batez ere, arretarik jarri gabe (edo gutxiago jarriz) agerkideen artean dagoen erlazio sintaktikoan, eta irizpide linguistikoetan oinarritutako kategoria edo sailkapenetan. Eragin handia izan du hizkuntzaren prozesamenduaren komunitatean, batez ere kolokazio-erauzketa automatikoaren hasierako lanetan (Church eta Hanks, 1990; Smadja, 1993) eta, ondorioz, teknika horietaz baliaturik eratu diren kolokazio-hiztegietan zein corpusak ustiatzen dituzten lexikografia-tresnetan.

Azken hamarkadan, hainbat ahalegin egin dira fraseologiaren aztergaiaren hedaduraz eta horren barnean sartuko liratekeen fenomeno linguistikoez, edo unitate-motez, aurkeztu diren ereduak bateratzeko edo hurbiltze-

ko, haien puntu komunak bilatzeko eta ikuspegi partekatuagoak proposatzeko. Jarraian, horietako batzuk aurkeztuko ditugu, gure lanaren markoa definitzen lagungarrien gertatu zaizkigunak, hain zuen ere.

Gries (2008)

Griesen iritziz, fraseologiaren definizio sendo batek sei parametro zehaztu behar lituzke. Hona hemen parametroak, eta hark bakoitzerako egiten duen proposamena. Griesek berak aitortzen duenez, oso ikuspegi zabaletik heldu dio parametro horiek zehazteko proposamenari, eskola desberdinetatik egin diren ekarpenak bildu nahian.

(i) the *nature* of the elements involved in a phraseologism

- elementuetako bat elementu lexikal baten forma edo lema izatea da baldintza, eta gainerakoak elementu lexikalak edo patroi gramatikalak izan daitezke

(ii) the *number* of elements involved in a phraseologism

- bi elementu edo gehiagoko unitateak kontsideratzen ditu

(iii) the *number of times* an expression must be observed before it counts as a phraseologism

- behatutako maiztasuna handiagoa izatea itxarondako maiztasuna, hau da, osagaiak ausaz konbinatuko balira itxaron litekeen maiztasuna baino

(iv) the permissible *distance* between the elements involved in a phraseologism

- elementuak ondoz ondokoak izatea ez du baldintzatzat jartzen

(v) the degree of *lexical and syntactic flexibility* of the elements involved

- espektro zabala onartzen du, mutur batean erabat zurrunak diren konbinazioak daudela (*by and large*), eta bestean, zehaztapen lexikal partziala duten konbinazioak (hala nola [$_{\text{VP}}$ DO into V-*ing*] moduko patroiak, non DO objektu zuzena den);

irizpide horrek kanpoan uzten dituen konbinazio bakarrak dira gutxienez elementu lexikal bat zehaztuta ez daukatenak (gramatika-patroiak, hala nola [$_{\text{VP}}$ V OBJ_1 OBJ_2])

(vi) the role that *semantic unity* and *semantic non-compositionality* / *non-predictability* play in the definition

- esanahi-unitatea izatea baldintza da, baina ez semantikoki konposizionala ez izatea

Beraz, Griesentzat:

> ≪A phraseologism is defined as the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance.≫

Sag et al. (2002); Baldwin eta Kim (2010)

Hizkuntzaren prozesamenduaren arloan izan duten eraginagatik, merezi du lan horiek aipatzea. Berez, ikuspegi estatistikotik abiatzen dira, *kolokazio* terminoa estatistikoki esanguratsua den edozein agerkidetza adierazteko erreserbatuz. Hala ere, HAUen prozesamenduari teknika estatistiko hutsez ez baina analisi linguistikoaren laguntzaz ekin behar zaiola argudiatzen dute.

HAUak definitzerakoan, Sag et al.-ek (2002) "MWEs are idiosyncratic interpretations that cross word boundaries (or spaces)" diote, eta hiru idiosinkrasia aipatzen dituzte: sintaktikoa, semantikoa eta estatistikoa. Idiosinkrasia da, beraz, konbinazio bat HAU egiten duen ezaugarri konposatua. Idiosinkrasia sintaktikoa malgutasunarekin dago erlazionatua; semantikoa, ez-konposizionaltasunarekin; eta estatistikoa, instituzionalizazioarekin, zehazki "maiztasun nabarmen handiz" agertzearekin. Idiosinkrasia-mota horien konbinazio-graduen arabera, HAU-motak bereizten dituzte.

Lan horretan oinarrituz, Baldwin eta Kimek (2010) HAUen definizio hau proposatzen dute:

> ≪Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity.≫

Terminologia-aldaketa nabari bat dago: idiosinkrasia gabe, *idiomatikotasun* erabili dute, eta honela definitu: lexema osagaien propietateetatik desbideratzea edo horiekiko markatua izatea[4].

Gainera, alderdi lexikala eta pragmatikoa gehitzen dizkiote Sag et al.-en (2002) idiosinkrasia sintaktiko, semantiko eta estatistikoen multzoari. Idiomatikotasun lexikala gertatzen da HAUaren osagai bat edo batzuk lexiko arruntaren parte ez direnean; pragmatikoa, berriz, HAUa testuinguru edo egoera jakin batekin edo batzuekin lotuta dagoenean. Bestetik, HAUaren osagaiak ondoz ondo ez agertzeko aukera kontuan hartzen dute.

Urizar (2012)

Azkenik, gure lanean ezinbesteko erreferentzia izan behar du euskarazko fraseologia konputazionalaren arloan egin den lan nagusiak, Ruben Urizarren *Euskal lokuzioen tratamendu konputazionala* tesiak. Lan horretan landutako marko teorikoan kokatu nahi izan dugu gure ikerketa, eta hurrengo lerroetan horren azalpen laburra egingo dugu. Bestetik, interesatzen zaigu Urizarren lana aurreko markoetan nola kokatzen den ikustea.

Urizarrek, Corpas Pastorren (1996) lanean oinarritu dela aitortuz, ezaugarri hauek landu ditu:

- **Polilexikalitatea**. UFak hitz batez baino gehiagoz osatuak dira. Hori zehazteko *hitz* terminoa bera zehaztu beharraz jabetuta, Urizarrek Linaresen (2006) definizioa darabil; horren arabera, *hitza* alderdi asko dituen unitatea da: a) forma foniko edo grafiko zehatza du; b) unitate gramatikala da; eta c) unitate lexiko-semantikoa da. Bestetik, Urizarrek ohartarazten gaitu hizkuntza idatziaren prozesamenduan ari denez hitz ortografikoez ari dela (zuriunez edo puntuazio-markaz bereiziak); gainera, euskara hizkuntza eranskaria izaki, UFak identifikatzeko osagaien lemekin ere lan egin behar da, ez forma flexionatuekin soilik.

- **Maiztasuna**. Urizarrek UFaren maiztasunaren bi alderdi nabarmentzen ditu. Batetik, UF baten osagaiak konbinaturik agertzen diren maiztasuna (*agerkidetza-maiztasuna*) handiagoa izaten da osagaiak ausaz konbinatuz gero espero litekeena baino. Bestetik, UFaren *erabilera-maiztasuna* zenbat eta handiagoa izan, aukera handiagoa dago esamolde finko gisa errotzeko; alderdi hori estuki loturik dago institucionalizazioarekin.

[4] "Idiomaticity refers to markedness or deviation from the basic properties of the component lexemes"

- **Instituzionalizazioa**. Lipka et al.-en (2004) azterketa aipatuz, instituzionalizazioa prozesu soziolinguistikoa dela dio Urizarrek, zeinen bidez "ale lexikal bat hiztun-komunitate baten norman integratzen da, haren hiztegiko lexema onargarri eta ohiko bihurtuz".

- **Egonkortasuna eta bariazioa**. Instituzionalizazioak erakusten dituen bi ezaugarri biltzen dira honetan: *finkapena* eta *espezializazio semantikoa*. Finkapena edo egonkortasun formala "alderdi lexiko-sintaktikoari dagozkion zenbait murriztapenen bitartez agertu ohi da". Bestetik, espezializazio semantikoa lotuta dago UFak jatorrizko interpretaziotik aldaketa semantiko bat izatearekin (*lexikalizazio* ere deitzen zaio horri). Aurreko bi alderdiok erlazionatuta daude, finkapen formalak aldaketa semantikoa ekarri ohi baitu. Bestetik, finkapenarekin lotutako alderdi garrantzitsu bat *aldagarritasuna* da, eta hor Urizarrek aldaera lexikalak (*begi bistan / begien bistan*) eta UFek diskurtsoan izan ditzaketen sormenezko aldakuntzak bereizten ditu.

- **Konposizionaltasunik eza**. Espezializazio semantikoaren edo lexikalizazioaren gradurik gorena dela esanez aurkezten du Urizarrek konposizionaltasunik eza. Aurrerago zehatzago ikusiko dugunez, konbinazioaren esanahia eta osagaien esanahien konbinazioa bat ez etortzea da. Puntu honetan aurkezten du Urizarrek *idiomatikotasun* terminoa. Kapitulu honen sarreran aurreratu dugunaren ildotik, terminoaren bi adiera bereizten ditu, bata ez-konposizionaltasunarekin identifikatua, eta bestea idiosinkrasiarekin erlazionatua, adiera zabalenean, hau da, UFen ezaugarri orokor edo metakontzeptu gisa.

- **Mailaketa**. UFek aurreko ezaugarri asko dituzte, baina maila desberdinean. Horren ondorioz, etengabeko *continuum* bat osatzen da.

Uste dugu aipagarria dela aurreko eskeman *instituzionalizazio* terminoa beste ezaugarri askorekin erlazionatuta aurkeztu izana. Batetik, UFaren erabilera-maiztasuna harekin estuki lotuta dagoela esan da; hurrena, finkapena eta espezializazio semantikoa instituzionalizazioak erakusten dituen bi ezaugarri direla; eta azkenik, ondoriozta liteke konposizionaltasunik eza, espezializazio semantikoaren goren gradua denez, instituzionalizazioak berekin dakarren zerbait ere badela. Beraz, balirudike instituzionalizazioa gertu dagoela gainerako ezaugarriak biltzen dituen ezaugarri orokor bat izatetik, ia UF-izaera edo "UFtasun" baten parekotik. Gainerako ezaugarri horiek, orduan, instituzionalizazioaren "sintomak" direla pentsa daiteke. Alde horretatik, instituzionalizazioa gertu legoke kapituluaren hasieran zirriborratu dugun idiomatikotasunaren kontzepzio zabaletik, metakontzeptutik.

Griesen (iv) parametroa izan ezik (unitate fraseologikoaren osagaien arteko testu-distantzia eta ondoz-ondokotasuna), gainerakoak ageri dira Corpas-Urizar ereduan. Bestetik, parametro edo ezaugarri bakoitzerako proposatzen diren irizpideak nahiko bat datoz, oro har. Aztertzen ari garen unitateek alderdi semantikotik, sintaktikotik zein estatistikotik halako berezitasunak dituztela aitortzen da. Baldwin eta Kimen (2010) idiomatikotasun pragmatikoaren ideia ez da Griesen ereduan ageri; Corpasenean, kontuan hartua dago, aurrerago ikusiko dugunez, UFetarako proposatzen duen sailkapenean[5].

Beharbada, desberdintasun nabariena da Griesen ereduan gramatika-patroiak izan daitezkeela UFen osagaiak edo elementuak. Aipatu dugun [$_{\text{VP}}$ DO into V-*ing*] eta antzeko patroiak fraseologiaren eremuan sartzeak eraman lezake bat pentsatzera, Griesek berak dioenez, "orain hizkuntzan dena fraseologikoa dela".

Fraseologiaren muga-arazoetako bat dago hor, hurrengo atalean beste batzuekin batera landuko duguna.

II.1.1.1 Fraseologiaren zenbait muga-arazo

González Reyk (1998) ohartarazten duenez, autore batzuek gramatika-elementuz (preposizioz edo bestelako partikulaz) osatutako konbinazioak (*of course, to give up, en pie...*) fraseologiaren eremutik kanpo uzten dituzte (Hausmann, 1989); eskola britainiarrean, berriz, kolokaziotzat hartu ohi dira (Benson et al., 1986). Zalantza horretaz ere dihardu Ruiz Gurillok (1998) *locuciones con casillas vacías* direlakoak azaltzean (*a juicio de*, *por parte de*; *en mi/tu/... caso*). Ruizen iritziz, horrelakoek izaera periferikoa dute fraseologian, ohiko konbinazioen eta lokuzio nuklearren arteko trantsizioa osatuz. Nolanahi ere, badirudi azkenaldiko proposamenetan, batez ere kolokazioen arloan egindako lanetan, ikuspegi inklusiboa nagusitzen ari dela (Gries, 2008: 5; Seretan, 2011: 25).

Horrek euskaraz ere inplikazioak ditu. Ez dago argi euskararen eta, oro har, edozein hizkuntza eranskariren kasuan, hitz anitzeko unitateen osaeran "hitz oso" ez diren elementuak, hots, morfemak, onar daitezkeen. Esaterako, zer dira *-z gero, -z batera, nahiz (eta) ... -n, -en aldean, -t(z)eko partez, -i utzi, -z baliatu*? Batzuk aditzen azpikategoria-sistemari dagozkio (*-i utzi, -z baliatu*), beste batzuk postposiziotzat jo daitezke (*-en aldean, -en bizkar*),

[5] Corpasen definizioan, nolabait inplizituki iragarrita dago hori, UFez dioenean "son unidades léxicas formadas por más de dos palabras gráficas en su límite inferior, cuyo límite superior se sitúa en el nivel de la oración compuesta."

eta beste batzuk menderagailu konplexuak lirateke (*nahiz eta... -n, harik eta... arte*).

Urizarren ereduan, bi kasutan onartzen da UFaren osagaiak morfema ez-independentea izatea (zehazki, kasu-atzizkia), lokuzio gramatikal kategoriaren barneko menderagailu-lokuzioetan eta postposizio-lokuzioetan (Urizar, 2012: 9-10).

Lokuzio lexikalen kategorian, ordea, hitz guztiak beregainak (edo hitz "osoak") dituzten unitateak baino ez ditu kontuan hartzen. Gure ikergaia kategoria horretakoa denez, irizpide horri lotuko gatzaizkio.

Beste bi muga-arazo daude: espezialitate-arloko hitz anitzeko terminoak, eta hitz anitzeko izendun entitateak. Gure ikergaiarekin lotura zuzena ez dutenez, ez ditugu hemen landuko, ideia orokor bat eman baino ez dugu egingo. Lehen kasuan, *ardo beltz* moduko unitate bat, ikuspegiaren edo aplikazioeremuaren arabera, kolokaziotzat har daiteke, edo kontzeptu zehatz bat adierazten duen terminotzat. Bigarren kasuan, irizpide orokorra da *Itsaso Horia* edo *Nazio Batuen Erakundea* ez direla fraseologia-arloko unitateak, onomastika-arlokoak baizik.

II.1.1.2 *Unitate fraseologiko* (UF) eta *hitz anitzeko unitate* (HAU) terminoak

Horiek horrela, zerbait esateko moduan gaude ikerlan honetan erabiliko dugun terminologiaz. Gure ikergaien izendapenerako, hiru aukera ditugu, lehen aipatutako *hitz anitzeko unitate* (HAU), *hitz anitzeko unitate lexikal* (HAUL) eta *unitate fraseologiko* (UF) terminoak.

HAU da kontzeptu zabalena, zeren HAUL terminoak, *stricto sensu*, ez baititu esaldi-izaera duten unitateak barnean hartzen. Egia da Ixa taldeak HAULen multzoan esaldi-erako unitateak ere sartzen dituela (Ezeiza, 2002: 96), baina, orain arte bederen, lexikalak soilik landu ditu. Bestetik, aurreko atalean ikusi dugu hitz anitzeko terminoak eta entitate-izenak badirela HAU, baina zalantzan dagoela UF ote diren. Beraz, HAU kontzeptua unitate fraseologiko kontzeptuaren hiperonimotzat jo dezakegu.

Tesi-lan honetan ikertu nahi ditugun unitateak UFen kategoriakoak dira, eta, HAU ere badira ere, termino zehatzena erabiltzea erabaki dugu. Irizpide hori bat dator Urizarrek bere lanean erabilitakoarekin (Urizar, 2012: 55). Izan ere, nazioartean ez ezik, Euskal Herriko espezialisten artean ere gero eta gehiago erabiltzen da.

II.1.2 Idiomatikotasunaren definizio operatiboa eta osagaiak

Aurreko atalean, ikusi dugu ezaugarri-multzo batek dakarrela hitz-konbinazio bat UF izatea, eta ez ezaugarri bakar batek. UF izate edo "UFtasun" delako hori izendatzerakoan, zenbait aukera daude: Sag et al.-en (2002) *idiosinkrasia*, Baldwin eta Kimen (2010) *idiomatikotasuna*, eta Corpas-Urizar ereduan gainerako ezaugarri biltzaile dela dirudien *instituzionalizazioa*. Beste proposamen bat *MWEhood* edo "HAUtasuna" da, *termhood* terminoarekiko analogiaz sortua (Baldwin, 2006; Hoang et al., 2009; Zaninello eta Nissim, 2010).

Azken urteotako fraseologia konputazionalean eta, bereziki, tesi honetan planteatu diren atazetan izan duen eraginagatik, *idiomatikotasun* terminoa da, gure kasuan, komenigarriena. Gainera, idiomatikotasuna "UFtasun"tzat kontsideratzea ez da HPren arloko berezitasuna. Ikusiko dugunez, fraseologiako hainbat autorek ere ikuspegi hori hartu dute.

Hala ere, esana dugu *idiomatikotasun* terminoa adiera desberdinez erabili dela fraseologiaren arloan. Unea da puntu hori sakonago lantzeko. Hauek dira idiomatikotasuna definitzeko erabili diren ikuspegi nagusiak:

- Idiomatikotasuna eta ez-konposizionaltasun semantikoa identifikatzen dituena. Hau da fraseologian tradizio handiena duena, eta, neurri batean, oraindik ere nagusitzat jo daitekeena (Gläser, 1998; Ruiz Gurillo, 1998: 19; Salvador, 2000: 19). Ez-konposizionaltasunaren ohiko definizioa da konbinazioaren esanahia ez dela osagaien esanahien konbinazioa, eta, beraz, ezin dela horien esanahiak konbinatuz eratu edo ulertu (Zuluaga, 1980: 123).

- Idiomatikotasuna UF izatearekin lotzen duena. Idiomatikotasunaren ideia irekitzen hasten da zenbait adituk esapide idiomatiko (*idiom, expresión idiomática...*) direlakoen propietateak esklusiboki semantikoak ez direla nabarmentzen dutenetik (Fernando eta Flavell, 1981; Fillmore et al., 1988; Barkema, 1996), propietate horiek gradualak direla (Ruiz Gurillo, 1998: 14), eta hitz anitzeko beste unitate-mota batzuetan ere aurkitzen direla, haien arteko continuum bat osatuz (Bolinger, 1977: 168, Cowie et al., 1983). Bada, Wulffek (2008) dio idiomatikotasunak, HAU guztien idiosinkrasiak atzeman nahi dituen terminoa izan nahi duenez, dagoeneko ez dituela konposizionaltasunik ezaren alderdiak soilik adierazten.

  Urizarrek (2012) aitortu duenez, idiomatikotasunaren zentzu zabal honek indar hartu du azken urteotan. Testuinguru honetan kokatu behar dugu lehen aurkeztutako Baldwin eta Kimen (2010) eredua.

- Idiomatikotasuna jatorrizko hiztunen adierazpen-hautaketarekin lotzen duena. Pawley eta Syderrek (1983) honela definitzen dute *native-like selection* terminoa:

  > ≪The ability of the native speaker routinely to convey his meaning by an expression that is not only grammatical but also nativelike; what is puzzling about this is how he selects a sentence that is natural and idiomatic from among the range of grammatically correct paraphrases, many of which are non-nativelike or highly marked usages.≫

  Ildo horretatik, Warrenek idiomatikotasunaren eredu bat landu du (Warren, 2005: 35-40), honela definitzen dena:

  > ≪Idiomaticity consists in knowing what situations and phenomena require standard expressions —although alternatives are normally conceivable— and in knowing what these would be.≫

  Idiomatikotasuna hiztunaren gaitasunaren osagai bat litzateke. Kontzepzio honen barnean sartzen dira aurrekoei egokitu dizkiegun alderdiak, baina zabalagoa da, diskurtso-egiturarekin eta pragmatikarekin erlazionatuta dauden gaitasunak dituelako.

  Ikuspegi horretatik oso gertu, Urizarrek aipatzen duen "partikulartasun" ideiarekin lotutako idiomatikotasuna dugu (Urizar, 2012: 66), hizkuntza jakin batek berezkoa eta berezia duena. Horretara, beste hizkuntza batekin alderatuta esan daiteke hizkuntza bateko esapideak idiomatikoak diren ala ez (Roberts, 1944), analisi kontrastiboa eginez.

Warrenen ereduak bere barnean hartzen ditu beste bi idiomatikotasun-kontzeptuen osagaiak, eta eredu zabalena dela esango genuke. Baldwin eta Kimek idiosinkrasia pragmatikoa sartu dute ereduan, eta Warrenen eredura hurbildu dira hein batean.

Gure ikergaia `izena+aditza` konbinazioetara mugatuta dagoenez, haren helmenaz haraindi daude, hein handi batean, Warrenen ereduan diskurtso-egiturari eta pragmatikari dagozkien osagaiak. Beraz, gure lanaren marko teoriko eta praktikorako, idiomatikotasunaren bigarren adiera hartu dugu UFen kontzeptua definitzeko eta sailkatzeko ezaugarri gakotzat, idiosinkrasia pragmatikoa alde batera utzita.

Horiek horrela, eta orain mahai gainean jarri ditugun ikuspegi guztiek eta beste hainbatek (Moon, 1998a: 6) partekatzen dituzten osagaiak instituzionalizazioa, konposizionaltasunik eza eta finkapena lexiko-sintaktikoa direnez, hau hartuko dugu idiomatikotasun definizio operatibotzat:

> Idiomatikotasuna konbinazio bat UF izatea determinatzen duen propietatea da, muineko ezaugarritzat idiosinkrasia duena (hizkuntzaren ohiko portaeratik aldentzea, banako hizkuntza-elementuen konbinazio libreak aurreikusi edo esplika ezin dezakeena). Idiomatikotasuna konplexua eta graduala da, eta bere barnean zenbait propietate hartzen ditu: instituzionalizazioa, ez-konposizionaltasun semantikoa (osoa edo partziala), eta finkapena (morfosintaktikoa zein lexikala).

Jarraian, definizio horretako propietate bakoitzean sakonago barneratuko gara.

II.1.2.1 Instituzionalizazioa

Urizarren lana aipatzean azaldu dugun definizioaren arabera[6], instituzionalizazioa prozesu soziolinguistiko bat da, zeinen bidez ale lexikal bat hiztun-komunitate baten norman integratzen baita, haren hiztegiko lexema onargarri eta ohiko bihurtuz (Bauer, 1983). Hiztunek unitatetzat hautematen dute konbinazioa, ezagutua eta aitortua da, eta, Corpas Pastorrek (2001) nabarmentzen duenez, konbinazioaren dimentsio psikolinguistikoa dugu hor, errealitate kognitiboa.

Aurrerago ere esana dugu instituzionalizazioaren sintomak UFen propietate edo idiosinkrasietan beha daitezkeela (semantikoa, lexiko-sintaktikoa, estatistikoa). Dena den, instituzionalizazioaren sintoma agerikoenetakotzat idiosinkrasia edo idiomatikotasun estatistikoa aipatu ohi da (Moon, 1998a: 7, Sag et al., 2002, Baldwin eta Kim, 2010: 7). Pecinak (2009) nabarmentzen duenez:

> «Institutionalized phrases, originally fully compositional and free word combinations, become significant and idiosyncratic by their frequent and consistent usage (especially in comparison with other alternative lexicalizations of the same concept).»

[6] Urizarrek ohartarazten gaitu termino honen inguruan dagoen nahaste kontzeptualaz (Urizar, 2012: 60), eta Lipka et al.-en (2004) azterketan oinarritu dela dio.

Beraz, osagaiek joera-maila bat agertzen dute elkarrekin konbinatzeko, eta emaitza ohikoa, ezaguna eta erabilia da. Ikerketa honetan, ikuspegi horretatik begiratuta erabiliko dugu *instituzionalizazio* terminoa.

Autore batzuek (Ruiz Gurillo, 1998: 20, Urizar, 2012: 60) konbinazioaren erabilera-maiztasunarekin lotzen dute instituzionalizazioa, eta ez, zehazki, agerkidetza-maiztasunarekin. Guk esango genuke, Griesen ildotik, estatistikoki idiosinkratikoa izatea ez dagoela zehatz-mehatz lotuta konbinazioaren maiztasunarekin, maiztasun hori osagaiak zori hutsez konbinatuko balira espero litekeen maiztasuna baino handiagoa izatearekin baizik, hau da, Urizarren agerkidetza-maiztasunarekin (Manning eta Schütze, 1999: 152).

Izan ere, berez maiztasun handikoak diren osagaien konbinazioak ere maiztasun handikoak izaten dira, baina zori hutsez gertatzen den fenomeno horrek ez lekarke, teorikoki behintzat, konbinazioa UFa izatea. Estatistikoki nabariak izaten dira, esaterako, aditz baten ohiko subjektuak edo objektuak dituzten konbinazioak: *liburua irakurri, ordenagailua piztu, ogia jan, aurpegia garbitu*... Ikuspegi horretatik, horrelakoak semantikoki motibatutako konbinazioak lirateke, hau da, konposizio libretzat hartzekoak (Bosque, 2001).

Bestetik, Baldwin eta Kimek (2010) zein Pecinak (2009) azpimarratzen dute idiomatikotasun estatistikoa, batez ere, konbinazioaren kontzeptu bera adierazteko lexikalizazio alternatiboekin konparatuta dela nabaria. Alderdi hori finkapen lexikalarekin dago, gure ustez, zuzenki lotua, eta dagokion atalean landuko dugu (II.1.2.3).

Azkenik, badirudi ideia intuitiboa dela konbinazio bat idiomatikoagoa izan ahala, idiomatikotasunaren propietate guztiak ere nabariagoak izango direla. Hala ere, idiomatikotasun estatistikoaren kasuan, beharbada gauzak ez dira hain ebidenteak, ez behintzat maiztasun hutsa kontuan hartzen bada. Izan ere, autore batzuek adierazi dute *idioms* direlakoak (lokuzioak edo esapide idiomatikoak), oro har, maiztasun txikiagokoak izaten direla kolokazioak baino (Moon, 1998b: 80), hau da, haien idiomatikotasuna ez dela berezitasun estatistikoaren ondorioa.

II.1.2.2 Ez-konposizionaltasun semantikoa

Konposizionaltasunaren printzipioa, Parteeren (1995) hitzetan, honela formula daiteke:

> ≪The meaning of a whole is a function of the meaning of the parts and of the way they are syntactically combined.≫

Hartara, hitz-konbinazio konposizional baten esanahia sintaktikoki konbinatuta dauden osagaien esanahien "konbinazioa" litzateke. Beraz, esapide bat idiomatikoa dela esaten da osagaien esanahien batura unitate edo konbinazioaren esanahiarekin bat ez datorrenean, konbinazioaren esanahia osagaien esanahietatik inferitu edo eratorri ezin daitekeenean (Cruse, 1986). Esaterako, *liburua irakurri* edo *autoa erosi* konposizionalak dira ('liburua irakurri' = 'liburua'+'irakurri'), baina *adarra jo* ez ('adarra jo' ≠ 'adarra' + 'jo').

Ez-konposizionaltasuna propietate bitartzat hartu izan da, esapide idiomatikoen eta gainerakoen arteko muga zehatz bat osatzeko balio duela argudiatuz (Makkai, 1972). Dena den, azkenaldian ugariagoak dira konposizionaltasun-mailak daudela diotenak, eta, konposizionaltasunik ezaz ez ezik, konposizionaltasun partzialaz ere mintzo direnak (Barkema, 1996: 140, Moon, 1998a: 34). Konposizionaltasun-maila horiek irizpidetzat erabili dira, aurrerago ikusiko dugunez, UFak sailkatzeko proposamenetan.

Eman dugun konposizionaltasunaren definizio sinplea aski intuitiboa bada ere, ez digu pentsarazi behar kontzeptu bakuna denik. Baditu ertz batzuk. Esate baterako, aurreko azalpenean aipatu dugun osagaien "esanahia" dela eta, ez da lehen begiratuan dirudien bezain sinplea. Nola zehaztu era objektibo batean zein den hitz baten "berezko" adiera, "oinarrizkoa", "arrunta" edo "prototipikoa"? Jatorrizkoa dela uler daiteke, edo ohikoena, eta horiek zein diren ere ez da beti gauza segurua.

Beste gai korapilatsu bat da osagaien esanahien konbinazioa nola eratzen den, nolako prozesua den. Ildo horretatik helduko gara konposizionaltasuna alderdi asko dituen ideia dela ikustera, beste zenbait kontzeptu gordetzen baititu barnean, edo berekin asoziatuta.

Izan ere, konposizionaltasunarekin erlazionatuta, maiz erabiltzen dira honelako terminoak: *motibazioa, analizagarritasuna, deskonposagarritasuna, gardentasuna/opakotasuna, esanahi literala/esanahi figuratiboa*... Denek dute zerikusia konbinazioaren esanahitik osagaien esanahien konbinaziora dagoen aldearekin, eta hau da, gure ustez, partekatzen duten ideia: konposizionaltzat jo ohi ez diren konbinazio batzuen esanahian, posible da nolabaiteko erlazio bat ezartzea konbinazioaren eta osagaien artean, haren esanahia motibatzeko, analizatzeko, deskonposatzeko edo ulertzeko bidea ematen duena. Urizarren hitzetan, "konbinazioaren dekonstrukzio semantiko bat gertatzen da, non esapidearen interpretazioaren zatiak osagai jakinekin uztartzen diren". Ezaugarri horri *deskoponsagarritasun semantiko* deritzo Nunberg et al.-en (1994) lanean. Hori ilustratzeko ingelesezko adibide aipatuenak dira *spill the beans* ('sekretua agerian utzi'; lit. 'indabak barreiatu') eta *kick the bucket* ('azkenak egin', 'akabatu', 'hil'; lit. 'baldea ostikatu').

Lehenean, esanahi ez-konposizionalean ere, izenaren eta aditzaren arteko banaketa bat dago; hartara, analogiazko asoziazio bat egin daiteke *spill* eta *reveal* artean, eta, batez ere esaldiaren egiturak ekarria bada ere (Svensson, 2008: 85), *beans* eta *secret* artean ere. Aldiz, ezin horrelakorik egin *kick the bucket* esapidearekin. Lehena deskonposagarria da, bigarrena ez. Euskarazko adibideak ematearren, esango genuke *zubiak eraiki* deskonposagarria dela, baina *adarra jo* ez.

Nolanahi ere den, Svenssonek (2008) ez ditu konposizionaltasunarekin erlazionatu ditugun aurreko termino horiek denak berdintzat ikusten, eta lau dikotomia bereizten ditu konposizionaltasunaren azterketan. Hauek dira dikotomiok, eta, oso labur, bakoitzari esleitzen dion bereizgarri nagusia:

- *motibazioa* vs *motibaziorik eza*: posible bada, behin esapidearen esanahia ikasitakoan, haren esanahia osagai bakoitzaren esanahietatik abiatuz azaltzea, esapidea motibatua da.

- *gardentasuna* vs *opakotasuna*: hiztunak esapidea arazorik gabe ulertzen badu, banako osagaien adierez gain aurretiko beste informaziorik gabe, esapidea gardena da.

- *analizagarritasuna* vs *analizaezintasuna*: esapidearen osagai bakoitzak haren esanahian duen ekarpena identifikatzea posible bada, esapidea analizagarria da. Deskonposagarritasunaren baliokidea da. Svesssonen arabera, analizagarritasuna ulertzeko modurik eraginkorrena da kontsideratzea egitura sintaktikoa bakarrik dela hura definitzen duena (egin berri dugun *spill the beans*en analisia da horren erakusgarri).

- *esanahi literala* vs *esanahi figuratiboa*: esapide baten hitzez hitzeko interpretazioa zentzugabea, absurdoa edo testuinguruarekiko desbideratua denean, figuratiboa edo metaforikoa da.

Lehen bi dikotomiak subjektiboagoak dira beste biak baino, eta zailagoak erabiltzen konposizionaltasuna ebazteko irizpidetzat. Bestetik, esapidearen arabera, dikotomia horiek modu desberdinean konbinatzen direla dio Svenssonek. Baina, horrez gain, ezaugarriak dikotomiatzat aurkezten dituen arren, Svenssonek aitortzen du partzialak ere izan daitezkeela, lehen adierazi dugunaren ildotik.

Horiek horrela, ezaugarri horien arteko mugek asko dute oraindik lausotik, ñabardura fin-finez bereiziak dira, eta eztabaidatzekoa izan daiteke idiomatikotasunaren karakterizazio automatikorako operatiboak diren, hau da, behar bezain behagai zehatzak antolatzeko aukera ematen duten.

II.1.2.3 Finkapena

UFen ezaugarri aipatuenetariko bat da ez dutela egitura bereko konbinazio libreen portaera lexiko-sintaktiko bera, zeren, horiek onartzen dituzten aldaera eta aldakuntzekin konparatuta, zurrunagoak baitira, hau da, zenbait murriztapen agertzen baitituzte (Ruiz Gurillo, 1998: 17, Manning eta Schütze, 1999: 173, Contreras eta Suñer, 2004: 90). Finkapena erabilerak eragindakoa da, arbitrarioa eta erlatiboa, hau da, murriztapenak zenbat eta handiagoak izan, UFa hainbat eta finkoagoa da, zurrunagoa eta, beraz, idiosinkratiko edo idiomatikoagoa; alderantziz, UFa malguagoa da, eta konbinazio libreetatik gertuago dago, murriztapenak gutxituz doazen neurrian[7].

Finkapena, *stricto sensu*, prozesu bat litzateke, zeinen bidez hitz-konbinazio bat konbinazio libreak baino zurrunagoa bihurtzen baita. Horren ondoriozko ezaugarria *finkotasuna* litzateke (Moon, 1998a; Fazly eta Stevenson, 2007). Finkotasunaren alderantzizkoa den *malgutasun* gisa deskribatzen dute autore batzuek ezaugarri hori (Barkema, 1994b; Bannard, 2007; Seretan, 2013).

Finkapenaren edo malgutasunaren barneko fenomenoak arakatzea eta bereiztea izan da ikertzaile askoren xedea, eta, oro har, adituak bat datoz kategoriak bereiztean. Batzuetan finkapenaren eskema orokor berean aurkezten diren arren, bi mota bereiztea komeni da, gure iritziz: morfosintaktikoa eta lexikala.

Finkapen morfosintaktikoa

Konbinazioak hizkuntzaren morfosintaxiaren araberako joko eta transformazioekiko duen malgutasun gisa definitzen da. Oso finkoak diren UFak ez dira erabiltzen egitura batzuetan, edo ez dute osagaien modifikatzailerik onartzen. Hein batean hizkuntzaren araberakoak izan daitezkeen arren, aldakuntza morfosintaktiko mota orokor batzuk zehatz daitezke. Zenbait ikertzaileren proposamenak aztertuta (Ruiz Gurillo, 1998: 17-19, Moon, 1998a: 104-174, Wulff, 2008: 87-143, Urizar, 2012: 101-106), hauek bereiz ditzakegu:

- Osagaien ordena-aldaketa

  Batzuetan, osagaien ordena aldaezina da: *gau eta egun* (eta ez **egun eta gau*).

- Osagaien flexioa

[7] Horrekin dago zuzenki erlazionatua II.1.1 atalean aurkeztu dugun aldagarritasuna (Urizar, 2012: 62-63), eta horren barruan bereizten diren aldaerak eta aldakuntzak.

Osagai batzuek ez dute edozein mugatzaile onartzen. Esaterako, *adarra jo* unitatean, *adar* izena singularrean da beti, eta *adarrak ipini* unitatean berriz, beti pluralean (dena den, biek onartzen dute partitiboa). Baina badira malgutasun mugatua dutenak; esaterako, *kontuan hartu* eta *kontutan hartu* erabiltzen dira[8].

- Osagaien hedapen sintagmatikoak edo modifikazioa

  Osagai batzuek ez dute ezer eranstea onartzen. *Adarra jo* unitateko *adar* izenari ezin zaio determinatzailerik, adjektiborik eta bestelako modifikatzailerik erantsi (**adar bat jo nion*, **adar ederra jo zenion*), eta ez du perpaus erlatiboetan erabiltzerik onartzen (**aurrekoan jo zenidan adarra ez zitzaidan gustatu*).

- Fokalizazioa eta topikalizazioa

  Zenbait esapidek ez dute honelako mugimendurik onartzen: **nork eman dio hitz?*; **gobernuak hartu du esku*. Dena den, gehienek onartzen dute, eta batzuek nahitaezkoa dute (*nork jo dizu adarra?* eta ez **nork adarra jo dizu?*).

- Gramatika-irregulartasunak

  UF batzuen osagaiak ez dira UFtik kanpo erabiltzen. Esaterako, *behinik behin* eta *fio izan* esapideetako *behinik* eta *fio*.

- UF batzuek sistematik kanpoko ezaugarri morfosintaktikoak dituzte: *lo egin*, *min eman* eta antzeko aditz-elkarteak irregularrak dira, objektu zuzen arruntak determinatzailea eskatzen baitu; sistemaren araberakoa *loa egin* litzateke (*ogia erosi* bezala, ez **ogi erosi*). Batzuetan, ordea, bi formak erabiltzen dira: *eztul/eztula egin*, *behar/beharra egin*.

Bada finkotasun morfosintaktikoaren eta konposizionaltasun-mailaren arteko erlazio bat. Oro har, konposizionaltasun apalak finkatze-maila handia ekarri ohi du berekin (Zabala, 2004: 465), eta, zehazkiago (Urizar, 2012: 100-103), konbinazioa deskonposagarria ez bada, ez du sormenezko aldakuntza morfosintaktikorik onartzen (Sag et al., 2002), eta bai, ordea, halakoa bada (Nunberg et al., 1994).

[8]Besterik da, noski, estandarrerako bakarra proposatzea, Euskaltzaindiak *Hiztegi Batu*an egin duena (*kontuan hartu*).

Finkapen lexikala

Konbinazioaren osagaiak (edo haietako bat, behintzat) ezin dira haien sinonimoez, kuasisinonimoez edo estuki erlazionatutako hitzez ordezkatu, emaitza ez delako erabiltzen, edo ez dituelako jatorrizko konbinazioaren ezaugarriak eta propietate idiomatikoak (Contreras eta Suñer, 2004: 95). Ezaugarri honi *kolokabilitate* (*collocability*) ere esan ohi zaio (Barkema, 1994b: 22-23), eta garrantzi handia du *kolokazio* kontzeptuaren definizioan (ikus II.2.2 atala)

Esaterako, *behin edo berriz* vs **behin edo berriro*; *babak eltzetik atera* vs **babak lapikotik atera*. `izena+aditza` konbinazioetara etorrita, *ziria sartu* esapidean ezin ditugu *zirir*en sinonimorik erabiliz **zotza sartu* edo **ezpala sartu* esapideak sortu. Ez dira existitzen. Bestetik, *erantzuna hartu* edo *erantzuna jaso* badira, baina *min hartu* da aukera bakarra, **min jaso*rik ez baita.

Murriztapen lexikala ere ez da beti erabatekoa eta argia izaten. Batzuetan, osagaiaren aukera lexikala itxia izan beharrean, sinonimo batzuk agertzen dira, baina, UF izango bada, aukera lexikal batekiko lehentasuna nabari behar da, ausaz legokiokeena baino handiagoa. Kolokazioetan gertatzen da batez ere. Esaterako, *urratsak egin* eta *pausoak eman* bide dira tradizioko testuetan gehien jaso diren konbinazioak, eta ez *urratsak eman* eta *pausoak egin*. Beraz, horrelakoetan, *murriztapen lexikal* terminoa baino zehatzagoak lirateke *lehentasun* edo *joera* terminoak.

UFen, eta batez ere kolokazioen, propietate honi garrantzi handia eman zaio hizkuntzen irakaskuntzan (Cowie eta Howarth, 1996; Keshavarz eta Salimi, 2007).

II.2 Idiomatikotasunaren continuuma eta UFen sailkapena

Idiomatikotasunaren definizioa landu ondoren, haren karakterizazioa zertan den argitzea da hurrengo egitekoa. Aurreko ataletan, adierazia dugu UF guztiak ez direla idiomatikotasun-maila berekoak, eta, gaingiroki bada ere, zenbait mota edo kategoria azaldu dira. Karakterizazioaren helburua litzateke, beraz, UF batek zer idiomatikotasun-maila duen zehaztea, edota zein kategoriatakoa den. Horretarako, UFen sailkapen-eredu bat aukeratu behar dugu lanerako.

UFen sailkapenari begira garrantzitsuak diren ideia gako batzuk izan behar ditugu gogoan.

Batetik, aurkeztu ditugun idiomatikotasunaren osagaiak hein desberdi-

nean konbinatzen dira, eta UFak sailkatzeko irizpideak ezartzearen emaitza graduazio-moduko continuum edo espektro jarraitu bat izaten da, argi bereizten diren kategoriak edo motak baino gehiago (Sinclair, 1996; Ruiz Gurillo, 1998; Bannard et al., 2003; Katz eta Giesbrecht, 2006; Wulff, 2008).

Bestetik, continuum horretan eremu edo zona desberdinak proposatu izan dira. Continuumaren errealitatea onartuta ere, badira arrazoi teorikoak zein praktikoak espektro jarraitu horretan zonak bereiz ditzakegula pentsatzeko. Azterketa sistematiko batzuk ere badaude, proposamenen arteko aldeak aztertu eta bateratze-saioak egin dituztenak. Hiru azterketa hauek izan ditugu kontuan gure azterketarako: Cowie (1998a: 4-8), Granger eta Paquot (2008: 35-44) eta Urizar (2012: 62-63).

Autore eta eskola gehienak bat datoz, terminologia desberdina erabiltzen badute ere, unitate fraseologikoen arteko lehen bereizketa honela egin daitekeela: unitate batzuek esaldi oso baten funtzioa bete dezakete, eta besteak sintagma-mailakoak dira (perpausetik beherako unitateak dira). II.1 taulan eman dugu lehen bereizketa hori egin duten zenbait autorek erabilitako terminologia.

- **Esaldi-unitateak**. Hizketa-egintza bat bete dezakete (Corpas Pastor, 1996: 269), eta definizio hori bat dator, gure ustez, *sentence-like unit* izendapena erabiltzen dutenek adierazi nahi dutenarekin (Gläser, 1998: 126). Enuntziatu fraseologiko ere esaten zaie. Horrelako unitateak dira, adibidez, *Egun on!, Eguzkia nora, zapiak hara* edo *Usteak erdi ustel*. Errutinazko formulak eta paremiak bereizi ohi dira, eta, horien barnean, atsotitzak, aipuak eta berariazko baliodun enuntziatuak (edo "esaldi eginak") (Urizar, 2012: 84-89).

- **Sintagma-unitateak**. Corpasen arabera, hizketa-egintza oso bat betetzen ez duten unitateak dira; Gläserren ereduan, *word-like unit* esaten zaie ("hitz-unitate"), edo *nominations*. Egokiago deritzogu, ordea, beste zenbaitek darabilten *sintagma-unitate* terminoari (Tristá Pérez, 1998: 300-301). Horrelakoak ditugu, adibidez, *hala eta guztiz ere, garunak urtu, min eman, erabakia hartu, gogo onez, adiskide min, izpi infragorri, arrain-sarda, lore-sorta*.

Gure ikergaiak sintagma-erako unitateen multzokoak direnez, horietan barnatuko gara. II.2 taulan, zenbait autorek sintagma-erako unitateak sailkatzeko egindako proposamenak daude.

II.2 taulako terminoen formak alde batera utzita, proposamenen arteko puntu komun bat da sintagma-unitateen barneko banaketa nagusia: esapide idiomatikoak (*idioms*, edo *locuciones*) eta kolokazioak. Aurreko taulen

Author	General category	Sentence-like (or pragmatic) unit	Word-like (or semantic) unit
Chernuisheva (1964)	Phraseological unit	Phraseological expression	—
Zgusta (1971)	Set combination	Set group	—
Mel'čuk (1988)	Phraseme, or Set phrase	Pragmatic phraseme, or Pragmategme	Semantic phraseme
Gläser (1988)	Phraseological unit	Proposition	Nomination
Cowie (1988)	Word-combination	Functional expression	Composite
Howarth (1996)	Word-combination	Functional expression	Composite unit
Corpas Pastor (1996)	Unidad fraseológica	Enunciado fraseolófico	No-enunciado fraseológico
Burger (1998)	Phraseological unit	Communicative PU	Referential PU

II.1 Taula: UFak sailkatzeko zenbait autoreren proposamenetan, *sentence-like unit* ("esaldi-unitate") eta *word-like unit* ("hitz-unitate") konbinazioetarako erabilitako terminoak. Zutabeen goiburuetan, Cowieren (1998a) bilduman erabilitako terminologia eman dugu.

Author	General category	Opaque, invariable unit	Partially motivated unit	Phraseologically bound unit
Vinogradov (1947)	Phraseological unit	Phraseological function	Phraseological unity	Phraseological communication
Amosova (1963)	Phraseological unit	Idiom	Idiom (nor differentiated)	Phraseme, or phraseoloid
Cowie (1981)	Composite	Pure idiom	Figurative idiom	Restricted collocation
Mel'čuk (1988)	Semantic phraseme	Idiom	Idiom (nor differentiated)	Collocation
Gläser (1988)	Nomination	Idiom	Idiom (nor differentiated)	Restricted collocation
Howarth (1996)	Composite unit	Pure idiom	Figurative idiom	Restricted collocation
Corpas Pastor (1996)	No-enunciado fraseológico	Locución	Locución	Colocación
Burger (1998)	Nominative	Idiom	Partial idiom	Collocation

II.2 Taula: UFak sailkatzeko zenbait autoreren proposamenetako *word-like unit* edo sintagma-unitateen azpikategoriak. Zutabeen goiburuetan, Cowieren (1998a) bilduman erabilitako terminologia eman dugu.

egituran antolatzen zaila den baina eragin handia izan duen beste tipologia bat Moonek proposatua da (Moon, 1998a: 19-25). Hiru makrokategoria bereziten ditu: a) *Formulak*, pragmatika-arazoekin lotuak (II.1 taulan aipatutako esaldi-motako unitateei dagozkienak); b) *Metaphors*, semantika-arazoekin lotuak (*idiom* motak biltzen dituztenak) eta c) *Anomalous collocations*, lexiko- eta gramatika-arazoekin lotuak (oro har, kolokazioak biltzen dituena). Makrokategoria bakoitzean zenbait kategoria zehazten ditu, baina onartzen du kategoria batzuk teilakatu egiten direla, eta UF batzuk kategoria batean baino gehiagotan sailka daitezkeela [Moonen (1998b) arabera, % 25].

Beraz, nahiko estandarra da idiomatikotasunaren continuumeko zona nabarienak esapide idiomatikoak eta kolokazioak direlako ikuspegia. Moonen hitzetan[9]:

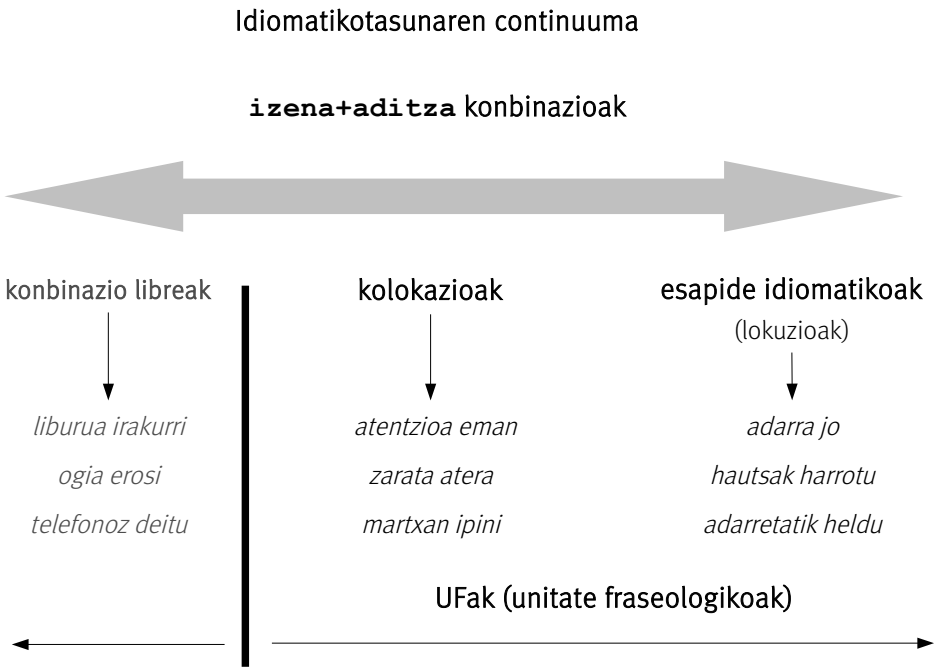
> ≪'Collocations' and 'idioms' represent two large and amorphous subgroups of FEIs on a continuum.≫

Horiek horrela, II.1 irudian ageri den eran irudika genezake `izena+aditza` osaerako sintagma-unitateen idiomatikotasunaren continuuma.

Ezker-muturrean, idiomatikotasuna 0 da; hau da, konbinazio libreak ez dira unitate fraseologikotzat hartzekoak. Eskuinekoan berriz, idiomatikotasuna 1 litzateke. Continuumaren eskualde horretan kokatzen ditugu esapide idiomatikoak (lokuzioak). Erdialdean, berriz, kolokazioak kokatu ohi dira. Horretara, kolokazioek, oro har, erdi-mailako intentsitatea agertuko luke te idiomatikotasunarekin erlazionatu ditugun fenomenoetan. Alde horretatik, esan genezake kolokazioak esapide "erdiidiomatikoak" direla (Polguère, 2000: 518).

Idiomatikotasunaren eta haren continuumaren kontzepzio hau nahiko koherentea da fraseologiaren ikuspegi kontestualistan eragin handia izan duen Sinclairren *idiom principle* ideiarekin (Sinclair, 1991: 110). Horren arabera, hizkuntza elkarren kontrako bi printzipiok gobernatzen dute: *open-choice principle* edo aukera libreko printzipioa, eta *idiom principle* edo idiomatikotasun-printzipioa. Lehenean, hitzak sistemaren arauen arabera konbinatzen dira, modu erregularrean, aukera guztiak eskura daudela; bigarrenean, konbinazio batzuk aurrez eratuta daude, bloke gisa erabiltzen dira, finkoak baitira. Bi printzipio horien arabera osatutako konbinazioak ardatz baten mutur banatan koka daitezke, II.1 irudian bezalatsu (Ooi eta Coi, 1998: 57; Kennedy, 1998: 109-110).

[9] FEIs: fixed expressions and idioms



II.1 Irudia: `izena+aditza` osaerako sintagma-unitateen idiomatikotasunaren continuuma.

Non da, ordea, esapide idiomatikoen eta kolokazioen arteko muga, edo zertan da bien arteko alde zehatza? Eta kolokazioen eta konbinazio libreen artekoa? Graduazio-kontu hutsa da, ala bada etenune gutxi-asko zehatzen bat mutur batetik bestera daraman continuum horretan?

Bestetik, II.2 taulako proposamenen arteko alde nabariena da autore batzuek esapide idiomatiko motak bereizten dituztela: *idiom* edo *pure idiom*ez gain, *partial idiom* edo *figurative idiom* bereizten dituzte. Deigarria da, gainera, zenbait sailkapenetan *collocation* eta *restricted collocation* terminoak agertzea; izan ere, *restricted collocation* terminoaren aurresuposizioa da *collocation* generiko bat badagoela. Orduan, bi kategoria nagusien azpikategoriak bereiz daitezke?

Hurrengo bi ataletan, esapide idiomatikoen eta kolokazioen ezaugarriak sakonago aztertuko ditugu, eta galdera horiei erantzuten ahaleginduko gara.

II.2.1 Esapide idiomatikoak

Esapide idiomatikoen edo lokuzioen taxonomia egiteko, irizpide desberdinak erabili izan dira.

Lehen irizpide bat funtzio sintaktikoaren araberakoa da. Esaterako, Urizarrek, Lorenteri (2001) jarraiki, lokuzio gramatikalak eta lexikalak bereizten ditu (Urizar, 2012: 107-109): "Lokuzio lexikalek hitz edo sintagma baten moduan funtzionatzen duten egiturak osatzen dituzte; lokuzio gramatikalak, ostera, unitate gramatikal gisa funtziona dezaketen estrukturak dira." Lokuzio lexikalak izen-, aditz-, adjektibo- adberbio- eta interjekzio-lokuzioak izan daitezke; gramatikalak, berriz, zenbatzaile-, izenordain-, loturazko eta postposizio-lokuzioak. II.1.1.1 atalean aurkeztu ditugun fraseologiaren muga-arazoetako bat lokuzio gramatikalekin lotua da, eta azaldu dugu, horren inguruan iritzi desberdinak dauden arren, gaur egun aski kategoria onartua dela.

Beste irizpide bat, berriz, idiomatikotasun-mailak bereiztea da. Finkapena eta konposizionaltasun semantikoa aztertuta, hainbat autorek adierazi dute esapide idiomatiko motak bereiz litezkeela. Idiomatikotasunaren karakterizazioaren aldetik, hau da interesatzen zaiguna.

Esaterako, finkapena kontuan izanik, begien bistakoa da esapide batzuk erabat finkoak direla, aldaezinak (Fraser, 1970; Moon, 1998a; Ruiz Gurillo, 1998). Adberbio-lokuzio eta lokuzio gramatikal ugari horrelakoak izaten dira. Esaterako, euskarazko *argi eta garbi, aldez edo moldez, hankaz gora, behinik behin*, eta beste hainbat halako. Bestelako esapideak, oro har, malguagoak izaten dira. Gure aztergai diren `izena+aditza` esapideak aski aldagarriak dira (Urizar, 2012: 129), eta VII.1.3.1 atalean, malgutasun morfosintaktikoa neurtzeko diseinu esperimentala egitean, zehatzago analizatuko dugu horrelako konbinazioen aldakuntzen gaia.

Konposizionaltasun semantikoa izan da, ordea, esapide idiomatikoen funtsezko izaera definitzeko eta azpikategoriak bereizteko gehien erabili den irizpidea. II.2 taularen iruzkinetan adierazi dugunez, autore batzuek esapide idiomatiko motak bereizten dituzte. Vinogradovek (1947) egin zuen estreinakoz bereizketa bat (*phraseological function* eta *phraseological unity*). Horren arabera, esapide idiomatiko batzuk erabat motibatu gabeak eta opakoak dira (*phraseological function* direlakoak), baina, beste batzuetan, motibazio partziala dago, esapide idiomatikoaren esanahia metafora bidez esplika edo interpreta litekeelako-edo. Gerora, Cowiek (1981) eta Howarthek (1998) bereizketa horri ongi iritzi eta beren sistemetan integratu dute, *pure idiom* eta *figurative idiom* izendapenak erabiliz.

Horrekin erlazionatuta, autore batzuek Makkairen (1972) bereizketaren

garrantzia nabarmentzen dute: esapide bat deskodetze aldetik izan daiteke idiomatikoa (*decoding idiom, idioms of decoding*), edota kodetze aldetik (*encoding idiom, idiom of encoding*). Zehazki (Fillmore et al., 1988: 504-505):

> ≪A decoding idiom is an expression which the language users couldn't interpret with complete confidence if they hadn't learned it separately. With an encoding idiom, by contrast, we have an expression which language users might or might not understand without prior experience, but concerning which they would not know that it is a conventional way of saying what it says. ≫

Deskodetze-esapideak opakoak direla uste dugu, eta kodetze-esapideen azpimultzo bat osatzen dute. Multzo horretan, opakoez gain, badira esapide figuratiboak eta murriztapen lexikal hutsa besterik ez duten bestelako UFak, hala nola kolokazioak. Euskarara etorrita, *adarra jo* deskodetze-esapidea dugu; *garunak urtu* eta *zubiak eraiki*, berriz, kodetze-esapideak.

Esapide idiomatikoen prozesamenduan hizkuntzaren erabilera figuratiboak eta, zehazki, metaforak zein metonimiak duten funtzioa arlo askotako ikergaia da; batez ere psikolinguistikarena eta hizkuntzalaritza kognitiboarena (Omazić, 2008), baina baita lexikografia zein fraseologiarena ere (Hanks, 2004). Zenbait teoria garatu dira esapide idiomatikoen interpretazioa esplikatzeko (Gibbs et al., 1997: 141-142). Oso arlo zabala da, eta gure ikergaiaz haraindi dago horietan barneratzea; dena den, interesatzen zaigu nabarmentzea maiz opakotzat jo ohi diren esapide idiomatikoak ere interpreta daitezkeela hizkuntzalaritza kognitiboaren ikuspegitik, hiztunen kontzeptu eta mekanismo kognitiboetan oinarrituta. Ez dugu ahaztu behar, hala ere, gure ikergaiaren helburu behinenetakoa dela erabiltzailearentzat deskodetzezein kodetze-arazoak dituzten esapideak erauztea eta karakterizatzea, hau da, hiztegigintzaren zein hizkuntzaren prozesamenduaren aldetik esplizituki deskribatzea merezi dutenak.

Fraseologiaren arlora etorrita, esapide idiomatiko opako eta figuratiboen arteko bereizketa erlazionatuta egon liteke II.1.2.2 atalean azaldu ditugun ez-konposizionaltasunaren alderdi batzuekin. Lehen begiratuan, esan genezake esapide idiomatiko opakoak ez direla deskonposagarriak, eta figuratiboak, aldiz, badirela, edo izan daitezkeela, behintzat. Izan ere, metaforaren mekanismoa esplikatzeko bide bat da osagaien esanahi literaletan oinarritutako adiera konposizionaletik nolabait abiatzea, eta horrek esan nahi du osagaien ekarpenaren aztarna hauteman dezakegula.

Baina ez dirudi gauzak hain sinpleak direnik. González Reyk Klinkenbergen (1990) metodoa erabili du metaforaren mekanismoa aztertzeko, eta

kasu hauek bereizten ditu esapide idiomatikoen artean (González Rey, 1998: 61-66):

- Konbinazio libre batetik eratorriak. Osagaiak semantikoki bateragarriak dira. Batzuetan, konbinazio librea arkaikoa da, edo erabilera urrikoa (*poner en la picota*); beste batzuetan, esapidea eta konbinazio librea bizirik daude egungo hizkeran (*echar leña al fuego*).

- Konbinazio libre batetik eratorri gabeak. Osagaiak ez dira semantikoki bateragarriak (*quemarse las pestañas, romperse la cabeza, llover a mares, echar un cigarro*).

Lehen kasuan, osagaiek ez dute identitatea galtzen, eta metaforaren mekanismoa abiarazteko bateraezintasuna konbinazio osoaren eta testuinguruaren artekoa da. *Echar leña al fuego* esapidearen kasuan, testuinguruaren arabera aktibatuko da esanahi literala edo figuratiboa. Bigarren kasuan, berriz, bateraezintasuna konbinazio barnekoa da, osagai baten edo batzuen erabilera metaforikoaren ondorioz sortua[10]. Ohar bat egin beharrean gaude hemen: hurrengo atalean ikusiko dugunez, autoreak semantikoki konposizionaltzat jotzen ditu kolokazioak, eta, horren ondorioz, aurreko sailkapenean ematen dituen adibide batzuk, zeinetan osagai batek bere ohiko edo oinarrizko adiera atxikitzen baitu, kolokaziotzat joko lituzkete beste zenbait autorek (esaterako, *llover a mares, echar un cigarro*).

González Reyren tipologiaren oinarrian, badirudi esapide idiomatiko guztietan metafora-mekanismo bat dagoela. Galdetzekoa litzateke, dena den, mekanismo horrek esapide erabat opakoak esplika ditzakeen. Izan ere, ez dugu uste *tomar el pelo* edo *bailar el agua* esapideak aurreko mekanismoen bidez azal daitezkeenik. Egokiagoa litzateke, segur aski, González Reyren sistemaren bidez azal daitezkeen esapide idiomatikoak figuratiboak direla esatea, eta, hortik kanpokoak, opakoak.

Euskarazko adibideak ematearren, badugu uste *zubiak eraiki* edo *ateak zabaldu* figuratibotzat har daitezkeela, baina zer esan *adarra jo*, *hanka sartu* eta *ziria sartu* esapideez? Badirudi horiek opakotzat jo behar genituzkeela. Bi kasu horien artean, badira oso zalantzazkoak eta, hiztunaren arabera, mota batekoak edo bestekoak izan daitezkeenak, hala nola *garunak urtu*,

[10]Gure ustez, deigarria da González Reyk *romperse la cabeza* konbinazioa kasu horretakoa dela esatea, zeren konbinazio librea bizirik baitago, eta horrek erakusten du ez dagoela osagaien arteko bateraezintasun semantikorik; esaterako: "estando de caza, se cayó del caballo, y se rompió la cabeza." Besterik da interpretazio literal horretatik metafora bidez irits gaitezkeen esapide idiomatikoaren esanahia motibatzera.

gerrikoa estutu eta *arantza atera*. Hiruko sailkapen hori batera samar letorke Moonek aipatzen dituen hiru metafora-moten sailkapenarekin: gardenak, erdigardenak eta opakoak (Moon, 1998a: 22-23).

Nolanahi ere den, Cowiek berak aitortzen du (Cowie, 1998b: 215) esapide idiomatiko opakoen eta figuratiboen arteko muga lausoa dela, eta faktore pertsonal, kultural edo idiosinkratikoen menpe egon daitekeela. Hiztun batek figuratibotzat jo dezakeen esapide bat erabat opakoa izan daiteke beste batentzat. González Reyk emandako adibideetara joz, esaterako, *poner en la picota* figuratibotzat hartu ahal izateko, jatorrizko konbinazio librea ezagutu behar genuke, eta hori norberaren kulturaren araberakoa izan daiteke. Areago, jakina da esapide askoren jatorria ezezaguna izaten dela, iluna edo badaezpadakoa, eta maiz espekulazio handia izaten dela tartean.

II.2.2 Kolokazioak

Kolokazioaren kontzeptua XX. mendearen bigarren erdian aurkeztu eta landu bada ere, lehendik ere zenbait adituk, batez ere lexikografia, fraseologia eta hizkuntza-irakaskuntzaren alorrekoek, aipatua zuten hitzen "konbinazioen" edo hitz bat beste hitz batekin edo batzuekin batera erabiltzeko "joeraren" ideia. Kennedyk aipatzen duenez, 250 urte iragan dira Alexander Crudenek Biblian hitz batzuk elkarrekin maiz agertzen direla ohartarazi zuenetik (Kennedy, 1998: 108).

Geroztik egindako lanen artean, Ballyrena da batik bat aipatzekoa (Corpas Pastor, 2001: 90). Hark azaldu zuenez (Bally, 1909), frantsesez *gravement* eta *griévement* sinonimotzat hartzea badago, baina *gravement malade* eta *grièvement blessé* erabiltzen dira, eta ez **gravement blessé* eta **griévement malade*. Horrelakoak hiztunentzat ohikoak edo ezagunak direla esan zuen, eta *groupements usuels* izendatu, trinkoagoak diren *groupes agglutinés* direlakoei kontrajarririk. Beste zenbait ikertzaile aritu dira gai honetaz; esaterako, Coseriu (1977), *solidaridad léxica* kontzeptuaren proposatzailea.

Dena den, Firthi aitortu ohi zaio *kolokazio* terminoa (*collocation*) estreinakoz erabili izana (Firth, 1957).

> ≪Collocations of a given word are statements of the habitual or customary places of that word.≫

Handik hona, kolokazioen inguruko gogoeta eta azterketak ugaritu egin dira, hiztegigintzan gero eta pisu handiagoa dute (Béjoint, 2000: 221-225; Evert, 2008: 2), hizkuntza-teknologien hainbat alorretan duten garrantzia

eta erabilgarritasuna gero eta argiagoa da (Manning eta Schütze, 1999; McKeown eta Radev, 2000; Seretan, 2013), eta hizkuntzaren teorian dihardutenen arreta ere erakarri du, batez ere, orain artekoan ikusi dugunez, fraseologiaren arloan. Azkenik, azpimarratzekoa da kolokazioak corpusen azterketaren eta corpus-hizkuntzalaritzaren testuinguruan landu direla bereziki, eta, batik bat, lexikografiaren eta testu-estatistiken alorrean.

Corpusak kontsultatzeko sistemetan, estandarra da hitz baten agerkidetzen informazioa ematea, eta lehen bide sinple bat da kolokazioak atzemateko. Bilatutako lemaren edo formaren, hau da, bilagaiaren, agerpenak konkordantzia-lerrotan edo KWIC (*Key Word in Context*) bistaratze-estiloan eskaintzeaz gain, bilagaiaren inguruan halako distantziara agertzen diren kategoria jakin bateko lemen zerrendak kontsulta daitezke. Horren adibide bat ipini dugu II.2 irudian: "*gol* (lema) + aditza" agerkidetzak.

II.2 Irudia: *Lexikoaren Behatokia* corpusaren kontsulta-sistemaren emaitzak "*gol* (lema) + aditza" agerkidetzetarako. Ezkerreko taulan, *gol* lemaren agerkide diren aditzak, maiztasunaren arabera. Horren eskuinean, agerpenen konkordantzia-lerroak edo KWIC (*Key Word in Context*).

Hiztegigintzan, aipagarria da, ohiko hiztegi orokorretan gero eta garrantzia handiagoa izan ez ezik, berariazko kolokazio-hiztegien garrantzia nabarmendu egin dela (Tutin, 2005), eta horrelako hainbat argitaratu direla zenbait hizkuntzatan. Horietako nagusien bilduma egin du Urizarrek bere tesian (Urizar, 2012: 92-93).

Euskaraz, berriz, oraindik ez dugu horrelakorik; izan ere, kolokazioena bide da euskaraz gutxien landu den fraseologia-arloa (Urizar, 2012: 81). II.3 atalean landuko dugu gai hori xeheago.

Kolokazioez mintzatzen denean, oinarrizko ezaugarri batzuk aipatu ohi

dira: kolokazioak "maiz gertatzen diren hitz-konbinazioak" dira, hiztunak "ohikotzat jotzen ditu, ezagunak dira harentzat", "ezin dira aurreikusi, arbitrarioak dira", "hizkuntzaren araberakoak dira" (*language-specific*), "hiztunak kolokazioak ezagutu eta erabiltzen jakin behar du hizkuntza batean jario naturala izateko", eta abar. Hori ilustratzeko, maiz adibideak aipatu ohi dira. Esaterako:

- en: *to pay attention, to draw conclusions, to score a goal; stiff breeze, big mistake, close friend, bitter enemy; bunch of flowers, flock of birds, shoal of fishes; completely wrong, extremely difficult*
- fr: *faire attention, tirer des conclusions, marquer un but; vent fort, grave erreur, amie intime/proche, ennemi déclaré/juré/né; bouquet de fleurs, volée d'oiseaux, banc de poissons; applaudir chaleureusement; difficilement compréhensible*
- es: *prestar atención, sacar conclusiones, meter/marcar un gol; fuerte viento, error garrafal, amigo íntimo, acérrimo enemigo; ramillete de flores, bandada/bando de pájaros, banco de peces; totalmente equivocado, extremadamente difícil*
- eu: *arreta ipini, ondorioak atera, gola sartu; haize zakar, akats larri, adiskide min, etsai amorratu; lore-sorta, txori-aldra, arrain-sarda; erabat oker, ikaragarri zail*

Horrelakoen berezitasuna zertan den eta bereizgarri komunik duten argitzea da jarraian egin behar duguna.

Firthek kolokazioaren kontzeptua aurkeztu zuenetik, kolokazioei eta hitz-konbinazioei buruzko ikerlan asko egin dira. Dena den, kontzeptuak eta haren aplikazioak halako garrantzia hartuta ere, ez da erabat lortu kolokazioaren definizio ahobatezko batera iristea, eta ikuspegi bat baino gehiago dago kolokazio kontzeptuaren hedadura zein den zedarritzeko (Tutin eta Grossmann, 2002; Evert, 2008: 1213 Seretan, 2011: 10-27).

Oro har (eta sinpletzeak berekin dakartzan arriskuak gorabehera), hurbilketa batzuek maiztasunean eta horrek berekin dakarren nabarmentasunaren neurketan jarri dute batez ere arreta; beste batzuetan, berriz, kolokazioaren osagaien arteko erlazioa ez da elkarrekin maiz agertzera mugatu, hori baino zerbait espezifikoagoa dela esaten da, eta, batzuen iritziz, propietate sintaktiko, semantiko eta konbinatorio bereziak dituzten agerkidetzak dira.

Horrenbestez, hainbat ikertzailek bi ikuspegi nagusi bereizi izan dituzte: estatistikoa eta linguistikoa (edo semantikoa). II.1.1 atalean aurkeztu

genituen bi ikuspegi horiek berak, unitate fraseologikoen testuinguru orokorrean. Han adierazi genuenez, oso ugariak dira bikoiztasun hori agerian jarri duten adituak. Euskarazko espezialisten artean, Corpasen eragina nabaria da (Altzibar, 2005: 2; Urizar, 2012: 95), eta horregatik ekarriko dugu hona haren ikuspegia. Corpasek dioenez (Corpas Pastor, 2001: 96-97) aipatu bi ikuspegiok kolokazioaren fenomenoa analizatzen eta ulertzen lagundu dute. *Kolokazio* kontzeptuaren bi alderdi hauek jasotzen ditu:

> ≪Entendemos por colocación aquella propiedad de las lenguas por la que los hablantes tienden a producir ciertas combinaciones de palabras de entre una cantidad de combinaciones teóricamente posibles (cf. Haensch et al., 1982: 251).≫

> ≪También denominaremos colocación a las combinaciones así resultantes, es decir, a las unidades fraseológicas formadas por dos unidades léxicas en relación sintáctica, que no constituyen, por sí mismas, actos de habla ni enunciados; y que, debido a su fijación en la norma, presentan restricciones de combinación establecidas por el uso, generalmente de base semántica: el colocado autónomo semánticamente (la base), no solo determina la elección del colocativo, sino que, además, selecciona en este una acepción especial, frecuentemente de carácter abstracto o figurativo (Corpas, 1996: 66).≫

Esango genuke, gainera, badela bien arteko lotura ageriko bat: murriztapen lexikala. Hain zuzen ere, posible diren konbinazio guztietatik hiztunek jakin batzuk erabiltzen badituzte, erabileran (hau da, norman) finkatzen da murriztapen lexikala. Beraz, hobe litzateke esatea kolokazioaren definizioan bi alderdi horiek aintzat hartu behar direla, hurbiltze kontrajarriak direla baieztatzea baino.

Beste hainbat autorek ere bi ikuspegiak integratzeko ideiarekin heldu diote kolokazioak definitzeari. Adibidez, Granger eta Paquotek (2008) Corpasen oso antzeko definizioa eman dute[11]. Baina batzuetan badira ñabardurak. Adibidez, Seretanek ezaugarri hauek lotzen dizkie: aurrefabrikatutako sintagmak dira, arbitrarioak, aurresanezinak, errekurrenteak, bi hitz

[11] "(Lexical) collocations are usage-determined or preferred syntagmatic relations between two lexemes in a specific syntactic pattern. Both lexemes make an isolable semantic contribution to the word combination but they do not have the same status. Semantically autonomous, the 'base' of a collocation is selected first by a language user for its independent meaning. The second element, i.e. the 'collocate' or 'collocator', is selected by and semantically dependent on the 'base'."

edo gehiagoz osatuak (Seretan, 2011: 14-17). Hau da, kolokazioak bi hitz baino gehiagokoak izan daitezkeela dio. Beste batzuetan, semantikan nabari dira aldeak: batzuen ustez, konposizionalak dira, eta, beste batzuek uste dute ezaugarri semantiko bereziak behintzat badituztela, partzialki ez-konposizionalak edo erdikonposizionalak izateraino.

Alderdi horiek aletzen joango gara jarraian.

II.2.2.1 Instituzionalizazioa

Corpasek aipatutako Haenschen definizioan, kolokazioekin maiz lotzen den "maiztasunaren" ideia da gunea, eta horixe izan da, beste hitzez bada ere, Firthen ideiari jarraituz kolokazioak lantzeko garatu den hurbilketa estatistikoaren oinarria, hau da, dagoeneko II.1.2.1 atalean instituzionalizazioaz mintzatu garenean aurkeztu dugun idiosinkrasia estatistikoa. Han adierazi dugu maiztasun absolutua, teorikoki bederen, ez dela nahiko irizpidea hori neurtzeko. Hain zuzen ere, corpusetatik kolokazioak erauzteko proposatu diren prozedura estatistikoen muinean, kolokazioen ideia hau dago: kolokazioa ausaz elkartuko liratekeen baino maizago batera agertzen diren hitzen multzoa da (McKeown eta Radev, 2000: 512).

Aurrerago, III.6.3 atalean, aurkeztuko ditugun neurri estatistikoak zoriaren arabera espero litekeenaren eta benetan behatzen denaren arteko konparazioan oinarritzen dira.

Baina, horrez gain, beste zehaztapen batzuk egin behar dira kolokazioaren kontzeptua zedarritzeko.

II.2.2.2 Polilexikalitatea: egitura eta osaera

Kolokazioen ikuspegi linguistikoaren lehen ekarria da kolokazioaren osagaien artean erlazio sintaktiko bat dagoela kontsideratzea. Ikuspegi sintaktikotik, autore batzuek kolokazioak bitan sailkatzen dituzte: kolokazio gramatikalak eta kolokazio lexikalak (Benson et al., 1986; Howarth, 1998; McKeown eta Radev, 2000: 511). Kolokazio gramatikaletan, hitz bat (izena, aditza, izenondoa) eta klase itxiko elementu bat (preposizio edo gramatika-egitura bat) konbinatzen dira[12]. Esaterako, ingelesezko *ready to, to be aware of, by accident*. Kolokazio lexikaletan, berriz, klase irekiko lexia beregainak dira osagaiak, eta ez dago preposizio, gramatika-egitura edo kidekorik.

[12] "Grammatical collocations are restricted combinations of a lexical and a grammatical word, typically verb/noun/adjective + preposition, e.g. *depend on, cope with, a contribution to, afraid of, angry at, interested in.*" (Granger eta Paquot, 2008: XX).

Sailkapen horren muinean esapide idiomatikoen atalean zirriborratu dugun funtzio sintaktikoaren araberako sailkapenaren ideia bera dago. De facto, aipatu egituren euskarazko baliokide izan litezkeen batzuk (postposizio bidez emanak denak ere) Urizarren lokuzio lexikalen sailean agertu dira (*-tik aitzin, -en arabera, -i begira, -z bestalde, -ik ezean, -en mende, -n zehar*). Beraz, II.1.1.1 atalean aurkeztu dugun muga-arazoa azaleratzen da hemen ere.

Kolokazio lexikalen egitura dela eta, bistakoa da egitura-sorta hizkuntzaren araberakoa dela. Euskaraz, Altzibarrek (Altzibar, 2005: 4-12) eta Urizarrek (Urizar, 2012: 99) egitura sintaktikoaren araberako taxonomia-proposamen bana egin dute, Corpasek gaztelaniarako emandako egitura sintaktiko sorta egokituta (Corpas Pastor, 1996: 270). II.3 eta II.4 irudietan eman ditugu.

Argitu beharra dago Altzibarrek,"Izena (objektua) + aditza" kategoriaren barnean, "adizlaguna (-n, -ra...) + aditza" egiturako kolokazioak aipatzen dituela, aparteko kategoria bat sortu gabe: *martxan jarri, eskura eman, kontuan izan*. Hark dioenez, horrelako hitz-konbinazioak batzuentzat aditz-perifrasiak dira, beste batzuentzat lokuzioak, eta beste batzuentzat kolokazioak.

Oro har, taxonomia horietan kontuan hartu diren egitura sintaktikoak nahiko bat datoz, sailkapena antolatzeko irizpideak desberdinak izan arren. Gure ikergaiarekin erlazionatuta dauden egiturak izen eta aditz batez osatuak dira, eta badira bi taxonomiaren arteko aldeak, baita eztabaidagai izan daitekeen egitura batzuen falta ere[13]. II.4 atalean aztertuko ditugu alderdi horiek, UFen testuinguru zabalagoan, eta IV.2.1 atalean diseinu esperimentalari dagozkien zehaztapenak.

Kolokazioen eredu sintaktikoak lantzea garrantzitsua da, batez ere, corpusetatik kolokazioak erauzteko estrategia bat eskaintzen duelako (hau da, kategoria jakin batzuetako hitzen konbinazioak soilik aztertzea). Corpusetatik kolokazioak erauzteko prozedura guztiek kontuan hartzen dute, era batera edo bestera, alderdi hori (ikus III.6).

Kolokazioen osagai-kopurua dela eta, teorian ez dago bi osagaiz baino gehiagoz osatuak izateko murriztapenik (Seretan, 2011: 16), eta definizio gehienetan, hala aitortzen da. Badira, dena den, itzal handiko autore batzuk, kolokazioak bi osagaiz osatuak direla diotenak (Hausmann, 1989; Mel'cuk eta Wanner, 1994), baina, funtsean fenomeno bitarra dela onartuta ere, errekurtsibitatea ere aitortzen zaio (Heid, 1994: 232): hau da, konbina-

[13] Esaterako, Gurrutxagak (2003) proposatutako taxonomian, `iz.(databoa)+ad.` egitura ageri da, *edanari eman, lanari ekin* modukoetan pentsatuta.

1 Izena dutenak oinarri

- 1. 1 Izena (objektua) + aditza: *hasiera eman, erabakia hartu, oztopoak jarri, bizimodua atera*
- 1. 2 Izena (subjektua) + aditza: *gerla piztu, zurrumurrua zabaldu; etsiak hartu, suak hartu*
- 1. 3 Izena + izenondoa: *adiskide zahar, harreman estu, gezur galant, istilu larri, esker on, etsai amorratu*
- 1. 4 Izenlaguna + izena (*-ko* markaren bidezkoa): *aldeko iritzi, erabateko lehentasun*
- 1. 5 Izena + izena: *baratxuri-atal, azukre-kozkor; arrain-sarda, txori-aldra*

2 Aditza dutenak oinarri

- 2. 1 Aditzondoa (modu- eta intentsitate-aditzondoa) + aditza: *argi esan, arlote jantzi, estu lotu, garbiki mintzatu, larriki zauritu*
- 2. 2 Onomatopeiaz osatuak (onomatopeiek aditzondo funtzioa betetzen baitute): *dzanga-dzanga edan, tipi-tapa joan, elurra mara-mara ari izan*

3 Izenondoa dutenak oinarri

- 3. 1 Izenondoa + izenondoa/partizipioa: *pobre erromes, txerri zikin; aberats okitu*
- 3. 2 Aditzondoa (maila-aditzondoa edo graduatzailea (*-ki*) + partizipioa (izenondo funtzioan)/ izenondoa: *itsuski kolpatua, larriki zauritua, erabat komentzitua*
- 3. 3 Adizlaguna (instrumental kasuan: *-z*) + partizipioa (izenondo funtzioan): *urguiluz hantua, jelosiaz urtua*

II.3 Irudia: Altzibarren kolokazioen taxonomia (Altzibar, 2005: 4-12).

1 Aditza + Izena (Subjektua): *haizea ibili, eguzkiak jo, zurrumurrua ibili*

2 Aditza + Izena (Objektua): *urratsa egin, izerdia bota, beldurra eman, erabakia hartu, erronka jo*

3 Aditza + Adberbioa:

3. 1 *gogotik/gogoz barre egin*

3. 2 *abian/martxan jarri, (korrika eta) presaka ibili*

4 Izena + Izena: *arrain sarda, euskal sen, giza baliabide*

5 Izena + Izenondoa: *astegun buruzuri, begi zoli, euskaldun peto, ezezko biribil, haserre bizi, kinka larri(an), lasaitu eder*

6 Izenlaguna + Izena: *ezinbesteko baldintza, gutxieneko soldata, bestelako kontu*

7 Adberbioa + Adjektiboa: *politikoki zuzen, guztiz bestelako*

II.4 Irudia: Urizarren kolokazioen taxonomia (Urizar, 2012: 99).

zio bitarraren bi osagaietako bat kolokazio bat ere izan daiteke (esaterako, *to play a central role, cometer un error garrafal*). Nolanahi ere, praktikan oso zabalduta dago bitarrak direlako ikuspegia, eta kolokazioak eraunzteko aplikazio gehienetan horri begira egin da lan.

II.2.2.3 Murrizketa lexikala eta konposizionaltasuna

Kolokazioei egokitu ohi zaizkien ezaugarri batzuk idiomatikotasun lexikalaren eta semantikoaren arlokoak direla uste dugu, eta, zehazki, murrizketa lexikalarekin dutela zerikusia. Hauek dira: *arbitrariotasuna, aurresanezintasuna, ordezkaezintasuna*, eta *polartasuna* edo *asimetria*. Elkarren artean erlazionatuta daude denak ere, jarraian ikusiko dugunez.

Funtsean dagoen fenomenoa da hitz batzuekin konbinatuta hitz jakin bat edo batzuk erabiltzen dituztela hiztunek, eta ez teorian erabil litezkeen beste batzuk, nahiz eta haien sinonimo edo baliokide izan. Beraz, *konbinazio-murrizketak* daude. Gainera, hitz jakin hori zein den ez dago modu

erregularrean aukeratzerik, *arbitrarioa* da; beraz, kodetzearen ikuspegitik kolokazioa *aurresanezina* da, eta osagai arbitrarioa, *ordezkaezina*. Deskodetzeari dagokionez, askotan konbinazioa ulergarria izaten da, aski gardena, baina arbitrarioa den hitzaren esanahia berezia izaten da, edo konbinazio horrekiko espezifikoa. Horregatik esan ohi da kolokazioak *polarrak* direla, osagai batek (*oinarria*) bere ohiko esanahia duelako, baina besteak ez (*kolokatiboa*). Hortik segitzen da kolokazioak, gehienetan, ez direla erabat konposizionalak, eta badutela idiomatikotasun semantiko maila bat. Azal ditzagun aurreko ezaugarriak xeheago.

Kolokazioak arbitrarioak direla esaten denean (McKeown eta Radev, 2000: 509-510; Tutin eta Grossmann, 2002: 7; Seretan, 2011: 15), adierazi nahi da kolokazio baten osagai bat ezin dela haren sinonimo batez ordezkatu, ez behintzat kolokazioaren izaera aldatu gabe, eta konbinazioaren propietate hori ezin dela hizkuntza-sistemaren arauetatik ondorioztatu, ez dagoela ageriko arrazoirik aukera bat edo bestea erabiltzeko. Esaterako, Manningek galdetzen du zergatik ingelesez erabiltzen den *stiff breeze* eta ez **stiff wind* (*light wind* erabiltzen da), *strong breeze* eta *strong wind* posible badira (Manning eta Schütze, 1999: 141). Horrelakoak dira, halaber, lehen eman ditugun Ballyren frantsesezko adibideak (*gravement malade* eta *grièvement blessé*), edo Corpasek aipatzen dituen gaztelaniazko *levar el ancla* eta *izar la bandera* (Corpas Pastor, 2001: 93).

Zuzeneko ondorio bat da ezin dugula aurreikusi zein den kolokazioan erabiltzen den sinonimoa eta zein ez, eta kolokazioa aurresanezina da. Beraz, kolokazioak erabileratik eskuratu edo ikasi behar dira. Ukaezina da horrelakoak deskribatzeak hizkuntza-irakaskuntzan eta itzulpenean duen garrantzia.

Arbitrarioa den osagaiak adiera berezia izaten du, edo konbinazioarekiko espezifikoa den adiera. Hausmannen (1989) arabera, kolokazioaren osagai batak, oinarriak, bere esanahia gordetzen du, eta, besteak, kolokatiboak, kolokaziotik kanpo ez duen esanahia hartzen du; horregatik dio kolokazioak konbinazio "polarrak" direla. Manningek adibide hauek aipatzen ditu, ikuspegi horren argigarri (Manning eta Schütze, 1999: 141): *strong tea* kolokazioan, *strong* izenondoak ez du 'indar fisiko handia duena' oinarrizko esanahia, 'agente aktibo asko duena' baizik; kolokazioko esanahiak badu nolabaiteko zerikusia oinarrizko esanahiarekin, baina ez da hura bera. Orobat *white wine, white hair* eta *white woman* kolokazioak, non *white* izenondoak kolore desberdinak adierazten baititu.

Bistan dena, konbinazioaren konposizionaltasunarekin zuzenean dago lotua hori dena. Azaldu dugun ikuspegi horretatik, kolokazioak erdikonposizionalak lirateke, eta hala aitortzen dute hainbat autorek (Mel'čuk, 1998:

30-31; Evert, 2005: 16; Seretan, 2011: 20-24).

Baina beste batzuentzat (Irsula, 1994: 277; González Rey, 1998: 60-61), kolokazioak konbinazio guztiz konposizionalak dira, eta, horregatik beharbada, fraseologia-eskola batzuek bere aztertze-eremutik kanpo edo periferian utzi dituzte (Corpas Pastor, 2001: 91). Ikuspegi horretan, kolokazioen eta konbinazio libreen arteko bereizgarria konbinazioaren osagaien arteko murrizte-erlazioren bat da, seguru aski erabiltze-maiztasunean islatzen dena; hori da, esaterako, Sag et al.-en (2002) ikuspegia sintagma instituzionalizatuak definitzerakoan.

Azkenik, badira bitariko kolokazioak daudela diotenak (Tutin eta Grossmann, 2002; McCarthy, 2008).

Horrenbestez, iritzi orokorrena kolokazioak erdikonposizionalak direla bada ere, batzuetan konposizionalak ere izan daitezkeela onartzea bide da irizpide zuhurrena. Hala izanik, murrizketa lexikala da, antza denez, kolokazio guztien propietate komuna.

Murrizketa lexikala

Azaldu berri ditugun propietate guztiak murrizketa lexikalarekin asoziatuta daude. Mel'čuken lana izan da kolokazioak edo, haren terminologian, agerkidetza lexikal murriztua (*restricted lexical cooccurrence*) ikuspegi semantikotik sistematikoki aztertzeko egin den saio nabarmenena. Honela definitu du (Mel'cuk eta Wanner, 1994: 325):

> ≪Restricted lexical cooccurrence is the cooccurrence of lexemes such that the choice of a specific lexeme L_1 for the expression of a given meaning is contingent on another lexeme L_2 to which the meaning is applied.≫

Mel'čuken ereduan, **AB** esapide idiomatikoaren **A** eta **B** osagaien 'A' eta 'B' adierazien batura ('A⊕B') ez da unitatearen esanahia; hori beste bat da ('C'). Kolokazioetan, berriz, ez da horrelakorik gertatzen. **A**, **B** eta **C**-ren artean dauden erlazio-moten arabera, Mel'čuk-ek zenbait kolokazio-kasu bereizten ditu. II.5 irudian eman dugu horien eskema.

1. kasuan, /A⊕B/ adierazlearen adierazia ez da 'A⊕B', 'A⊕C' baizik. Beraz, unitatearen osagai baten adierazlearen (/A/) adierazia erregularra da ('A'), baina bestearena ez. 1.(a) kasuan, **B** lexema eduki semantikorik gabea da, edo semantikoki oso murriztua dago; 1.(b) kasuan, **A**-rekin konbinatzen denean hartzen du **B** lexemak 'C' esanahia. Horrelakoetan, kolokazioa erdikonposizionala da. Aipatutako euskarri-aditz edo aditz arin direlakoak 'A'

1 'C'≠'B' (**B**-k ez du, hiztegian, 'C' esanahia)

(a) 'C' hutsa da, hau da, **B** euskarri-aditza edo aditz arina da: *to do a favour, to give a look, to take a step*

(b) 'C' ez da hutsa, baina **B** lexemak **A**-rekin (edo kideko lexema-multzo mugatu batekin) konbinatuta soilik du 'C' esanahia: *black coffee, French window*

2 'C'='B' (**B**-k badu, hiztegian, 'C' esanahia)

(a) 'B' ezin da **A**-rekin konbinazioan adierazi **B**-ren sinonimo baten bidez: *strong (*powerful) coffee, heavy (*weighty) smoker*

(b) 'B'-ren barnean 'A' adierazlearen parte bat dago, hau da, **B** lexema **A**-rekin lotua da: *artesian well, aquiline nose*

II.5 Irudia: Mel'čuken kolokazioen taxonomia (Mel'čuk, 1998: 30-31).

adieraziari dagokion egoera edo ekintza adierazteko erabiltzen diren eduki semantiko gutxiko aditzak dira, balio predikatibo gutxikoak. Euskarazko `izena+aditza` osaerako UFak lantzean jardungo dugu horietaz sakonago (II.4 atala).

2. kasuan, kolokazioa konposizionaltzat jotzen da. 2.a kasuan, murrizketa lexikal hutsa gertatzen da; 2.b kasuan, kolokaziokidearen esanahia /A/ adierazlearekin erabat asoziatua dago (**B** is "bound" to **A**).

Mel'čukek bere Meaning↔Text teoriaren barnean egin du hitzen konbinazioen sailkapen semantikoa (Mel'čuk eta Polguere, 1987). Teoria hori kolokazioen sailkapen semantikoaz haraindikoa da, hizkuntzaren teoria bat da, baina, eskuartean dugun aztergaiari gagozkiola, muineko kontzeptua funtzio lexikala da (LF). Funtzio lexikalak bi hitzen arteko erlazio semantiko-sintaktiko orokor bat adierazten du. LFen bidez, kolokazioaren oinarria (L_1) eta kolokatiboa (L_2) honela erlazionatzen dira: $f(L_1) = L_2$. L_1 hitzari *gako-hitz* deritzo, eta, L_2-ri, *balio*. Esaterako, II.3 taulan, Alonsok emandako adibide

batzuk daude (Alonso, 1996: 53).

Funtzioa (gako-hitza)	**balioa**
Magn(*deseo*)	*ardiente*
Magn(*ganas*)	*locas*
Magn(*probihir*)	*terminantemente*
$\mathbf{Oper}_1$(*beso*)	*dar* [*un* ∼]
$\mathbf{Oper}_1$(*caricia*)	*hacer* [*una* ∼]

II.3 Taula: Funtzio lexikalen eta haien balioen zenbait adibide (Alonso, 1996: 53).

Magn eta $\mathbf{Oper}_1$ funtzio lexikalak dira, kolokazioaren osagaien arteko erlazioa adierazten duten eragile orokorrak, alegia. Hainbat funtzio lexikal deskribatu dira (60 inguru) eta jadanik horietako batzuk zenbait lanetan aplikatzen hasi dira kolokazioen sistematizazioa egiteko; esaterako, DECIDE proiektuan (Grefenstette et al., 1996), eta *Explanatory Conbinatorial Dictionary* erako hiztegi-proiektuetan, hala nola *Dictionnaire explicatif et combinatoire du français contemporain* (Mel'čuk et al., 1984; Mel'čuk eta Polguère, 2007) eta *Diccionario de Colocaciones del Español* proiektuan[14] (Alonso, 2004).

Mel'čuken sisteman, kolokazio guztiak dira lexikalki murriztuak, erdikonposizionalak zein konposizionalak izan. Dena den, autore batzuek erakutsi dute murrizketa hori batzuetan ez dela esklusiboa (Contreras eta Suñer, 2004: 54). Batetik, kolokatibo bat antzeko adierarekin ager daiteke kolokazio batzuetan, oinarri desberdinekin konbinatuta: *zarata/istilua/iskanbila... atera* (Urizar, 2012: 96). Bestetik, oinarri berarekin, antzeko adiera duten kolokatibo batzuk erabil daitezke, hala nola *apagar/saciar/matar/satisfacer la sed* (Koike, 1998: 246).

Horrek garamatza "kolokazio irekiak" direlakoen auzira. Aurreko agerkidetza-motak azaltzeko, autore batzuek (Aisenstadt, 1981; Cowie, 1986; Howarth, 1996) *kolokazio ireki* edo *kolokazio libre* kontzeptua erabiltzen dute (*free/open collocation*), *restricted collocation*etik bereiziz, hitz bat sorta bateko item lexikalekin konbinatzen dela adierazteko. Esaterako, *hire* aditzak hitz ugarirekin osa ditzake ohiko konbinazioak: *staff, clerk, secretary, worker*. Horrelakoei *kolokazio semantikoak* (Viegas eta Bouillon, 1994), edo "motibatuak" (Mel'cuk eta Wanner, 1994: 325) deitu izan zaie (Mel'čuken

[14] http://www.dicesp.com/

hitzetan, "collocates may correlate with the semantic of the base"). Kontzeptua oso eztabaidatua izan da, batez ere horrelako agerkidetza murriztuak hautatze-murrizketen bidez esplika daitezkeelako, kolokazio-murrizketen bidez baino egokiago (Fontenelle, 1998: 192). Fraseologiaren muga-arazo nabari bat da hori.

II.2.2.4 Malgutasun morfosintaktikoa

Eskuarki, kolokazioak morfosintaktikoki murrizketarik gabeak edo oso gutxikoak direla esan ohi da, hau da, konbinazio libreen malgutasun-maila bertsukoak (Urizar, 2012: 97; Vincze, 2013: 244).

Hala ere, malgutasun hori oso aldakorra dela, eta, batzuetan, aski mugatua izan daitekeela ere adierazi da. Esaterako, konbinazioaren egitura sintaktikoak zerikusia bide du malgutasunean (Contreras eta Suñer, 2004: 53): gaztelaniazko `izena+izenondoa` kolokazioetan, eragiketa sintaktiko batzuek naturalak ez diren aldakuntzak sor ditzakete. Horren antzeko euskarazko kasuak ditugu, esaterako, *adiskide min* edo *lagun hurko* modukoak, kolokazio argiak direnak baina honela nekez erabiltzen direnak: *nire adiskide handia da, izan dudan minena*. Bestetik, frantsesezko aditz arinen eraikuntza batzuk modu atipiko edo arkaikoan eratuak dira (Tutin eta Grossmann, 2002: 8-9), izena determinatzailerik gabe gerta baitaiteke (*avoir faim, rendre visite*), gaur egun estandarra denaren kontra (*perpétrer un délit*). Berezitasun horrek euskarazko konbinazio asko dakartza gogora (*lan egin, min hartu*). Euskarazko `izena+aditza` osaerako konbinazioak berariaz landuko ditugu II.4 atalean.

Nolanahi ere, kolokazioen malgutasun morfosintaktikoa konbinazio libreen neurri berekoa ez dela onartuta ere, diskriminatze-ahalmen txikiko propietatea dela aitortu ohi da (Heid, 1994: 232).

Laburbilduz, atributu hauek osatzen dute ikerkuntza honetan erabiliko dugun *kolokazio* kontzeptua:

- Hitz-konbinazio egonkorrak dira, ohikoak eta instituzionalizatuak, norman finkatuak (eta ez sisteman).
- Erlazio sintaktikoa duten bi elementuz osatuak dira, baina errekurtsibitatea agertzen dute, eta osagai gehiagoko konbinazioak osa daitezke.
- Konbinazio-murriztapena agertzen dute (murriztapen lexikala), bi eratara:

  - osagai batek (oinarriak) bere esanahi ohiko edo arrunta atxikitzen du, eta besteak (kolokatiboak) eduki semantiko gutxiko edo konbinazioarekiko espezifikoa den adiera bat hartzen du; konbinazioa erdikonposizionala da.
  - osagaiek esanahi ohiko edo arrunta atxikitzen dute, baina haietako bat (kolokatiboa) ezin da sinonimoz edo baliokidez ordezkatu; konbinazioa konposizionala da.

    Batean zein bestean, oinarriaren adiera ohikoa da, eta horrek bereizten ditu kolokazioak esapide idiomatikoetatik.

- Morfosintaktikoki, ez dute konbinazio libreen malgutasun bera, batzuetan murrizketak agertzen baitituzte, baina ez da propietate bereizgarria.

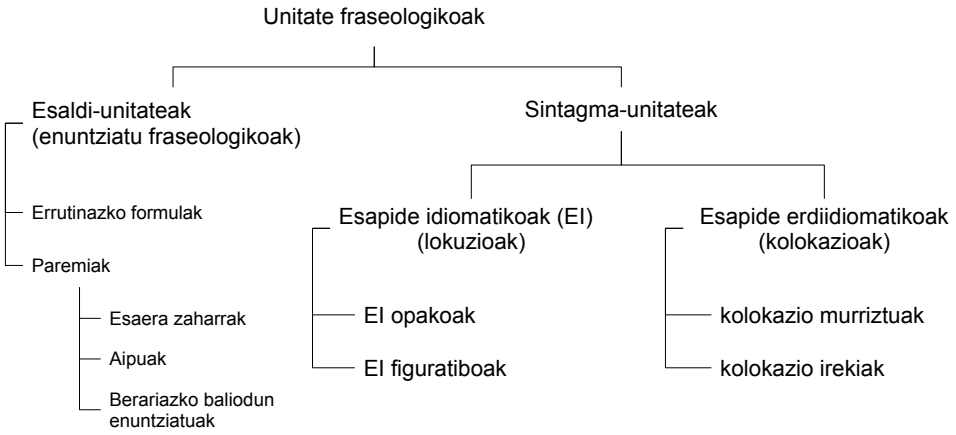
II.2.3 Sailkapen-proposamena

Aurreko ataletan, sintagma-erako bi UF kategoria nagusien (esapide idiomatikoen eta kolokazioen) ezaugarriak landu ditugu, eta bakoitzaren barnean proposatu diren azpikategoriak azaldu. Ikerkuntza honen helburua `izena+aditza` osaerako UFen karakterizazioa denez, sailkapen-proposamen bat aukeratu behar da atazaren jomugatzat.

Orain arte aztertu ditugun proposamenak kontuan izanik, II.6 irudiko sailkapena proposatzen dugu. Gure ikergai ez diren arren, esaldi-unitateak ere kontuan hartu ditugu.

Sailkapen horretako kategoriak eta azpikategoriak Cowie (1986) eta Howarth (1996) lanetan oinarrituak dira. Bat datoz Ezeizaren doktorego-tesian proposatutakoarekin (Ezeiza, 2002: 96-97), nahiz eta azpikategorien definizioak eta hedadurak ez diren erabat baliokideak.

UFak bi kategoria nagusitan banatuta daude: esapide idiomatikoak (edo lokuzioak) eta erdiidiomatikoak (edo kolokazioak). *Esapide erdiidiomatiko* terminoa sartu dugu eskeman, batetik, laguntzen duelako, *esapide idiomatiko* terminoaren ondoan, idiomatikotasun-continuumeko bi zona nagusien ideia egoki jasotzen. Beste arrazoi bat hau da: aurrerago ikusiko dugunez, badira normalean kolokazio kontzeptuarekin lotzen ez diren baina horiekin ezaugarri garrantzitsuak partekatzen dituzten UF batzuk (ikus II.4 atala), eta horiek continuum horretan kokatzeko aukera egokia eman lezake *esapide*



II.6 Irudia: UFen sailkapen-eredua.

erdiidiomatiko terminoak. Nolanahi ere, hain da *kolokazio* terminoa erabilia eta zabaldua, ezin baitzaio erabiltzeari utzi.

Ikusi dugu azpikategorien arteko mugak lausoak direla, eta, kolokazioen kasuan, proposatutako azpikategoria bat (kolokazio irekia), aski eztabaidatua dela, kuestionatu baita fraseologiaren arloko fenomenoa den. Beraz, proposamen honen bideragarritasuna hizkuntzalari batzuek egingo duten eskuzko sailkatze-lanaren bidezko egiaztatze eta balidatze baten mende geratzen da (ikus VI.4).

II.3 Euskarazko fraseologiaren ikuspegi laburra

Euskarazko fraseologian egin den lana gaingiroki azalduko dugu atal honetan, gure ikerkuntza zein testuingurutan txertatzen den argiago ikusaraztearren. Arlo honen ikuspegi orokorra jaso duten lan nagusiak Kaltzakorta (2001), Esnal (2001), Urizar (2012) eta Altzibarrenak (2013) dira. Hauek dira, laburbildurik, Urizarrek aurkezten dituen euskarazko fraseologismo-bilduma nagusiak eta haien ezaugarriak:

- *Refranes y Sentencias*, 1596koa, egile ezezagunekoa, Lakarra (1996) lanean argitaratuak; 273 esaera zahar.
- Esteban Garibaik (1533-1599) aipatutako esaera zaharren bi bilduma (guztira, 103).
- Arnaud Oihenarten 1657ko atsotitz-bilduma (Altuna eta Casares, 2003).

- Pablo Zamarriparen *Manual del Vascófilo* liburuan argitaratutako euskal lokuzioen bilduma (Zamarripa, 1913).

- Resurrección María Azkueren *Atsotitzak / Proverbios* eta *Esakerak / Modismos* bildumak, 1935-1947 bitartean *Euskalerriaren Yakintza* izenpean argitaratuak (Azkue, 1989). Lehenak 3 000 esaera zahar inguru ditu, gaika sailkatuak. Herenak baino gutxiago dira liburuetatik jasoak; gainerako guztiak, hiztunengandik. Bigarrena lokuzio-bilduma da. Gehienak hiztunengandik jasoak dira (1 700), eta gainerakoak idazleen eta esapide-biltzaileen lanetatik (311).

- Jean Elizaldek *Gure Herria* aldizkarian 1 200 bat "zuhur hitz eta erran zahar" argitaratu zituen, 1936tik aurrera. Landa-lanaren bidez eta Oihenarten liburutik hartuak.

- Paul Gilsouren *Errantegia*. 382 orrialdeko lan argitaragabea da, 1964an burutua (Kaltzakorta, 2001: 84). Paremiez gain, lokuzio batzuk ere biltzen ditu.

- Damaso Intzaren *Naparroa-ko euskal-esaera zarrak* bilduma, bi bilketa-aldiren emaitza dena. 2 500 "esaera" jasotzen ditu, eskualdeka sailkatuak. Gehientsuenak atsotitzak badira ere, lokuzioek leku berezia dute bertan.

- Antonio Zabalaren *Esaera zaarren bilduma berria* izeneko bi liburukiak (Zavala, 1985). 3 133 esaera.

- Gotzon Garateren *Atsotitzak* bilduma (Garate, 2003). 30 000 esaera zahar inguru, Euskal Herri osokoak. Gehienak landa-lanean bilduak; batzuk beste bildumetatik hartuak, edo laguntzaileengandik. Interneten kontsulta daiteke[15]. Eleaniztuna da: euskara (14 458), gaztelania (5 208), ingelesa (4 045) eta latina (3 462).

- Koldo Izagirreren *Euskal lokuzioak* lana (Izagirre, 1981). Auspoa argitaletxeko liburuak hartzen ditu bere lanaren corpus nagusi moduan, bertsolarien jarduna aho-hizkeratik hurbilena omen delako. 7 000 lokuzio inguru jasotzen ditu, eta erdiak inguru aditz-lokuzioei dagozkie. Lan honetan oinarrituta, *Intza proiektua* abiatu zen, *Euskal Lokuzioak Sarean* webgunean[16]. Geroztik, lana osatu eta orraztu egin dute.

[15] http://www.ametza.com/bbk/htdocs/garate.htm
[16] http://intza.armiarma.com/

- Justo Maria Mokoroaren *Ortik eta hemendik. Repertorio de locuciones del habla popular vasca, oral y escrita, en sus diversas variedades* bilduma (Mokoroa, 1990). 92 000 esaldi edo adibidetik gora biltzen ditu. Informazio-iturri zabala erabili da: idazleak, egunkariak eta solasaldiak; guztira, 385 iturburu bibliografiko eta 1 300 bat lekuko. Lokuzioak ez ezik, kolokazioak eta enuntziatu fraseologikoak ere biltzen ditu (Esnal, 2001: 140). Online kontsulta daiteke[17].

Lehenik eta behin, aitortu beharra dago ia ez dela teorizazio-lanik edo azterketa teorikorik egin euskal fraseologiaren gainean (Kaltzakorta, 2001: 79).

Eratutako bildumak direla eta, Urizarrek (2012) nabarmentzen du luzea eta oparoa dela euskarazko atsotitz edo esaera zaharren zein, neurri txikiagoan bada ere, lokuzioen bildumen ibilbidea, oso bilduma handiak osatzeraino, maiz gaika sailkatuta daudenak, eta, zenbaitetan, erdarekiko baliokidetzak ere eskaintzen dituztenak. Gainera, ikusi dugu bilduma nagusi batzuk online kontsultatzeko aukera ere badagoela. Horrek eragin handia du, esan beharrik ez dago, material horien zabalkundean eta gizarteratzean.

Bilduma horietan jasotako UF-kategoriak direla eta, bada aipatzea merezi duen gai bat: hizkuntza-jarduera ohikoan edo egunerokoan, baliagarritasun gutxiagokoak izaten dira paremiak lokuzioak baino (eta seguru asko, lokuzioak ere erabilera txikiagokoak kolokazioak baino). Hala ere, horiexek erakarri dute espezialisten arreta. Litekeena da horren arrazoia izatea esaera zahar eta atsotitzez dugun irudia edo ikuspegia, herri-kulturaren agergarri gisa, eta horiei garrantzi handia eman ohi zaie, batez ere diglosia edo mendekotasunean bizi den herri eta kultura batean. Hain zuen ere, aurreko lan askoren egileen kezka handienetako bat da horrelako hizkuntza-baliabideak galbidean direla, gero eta gutxiago entzuten direla, eta desagertzeko zorian daudela. Oso nabaria da hori, batez ere, Mokoroarengan eta Garaterengan. Horrek are beharrezkoagoa egiten du, jakina, lehenbailehen biltzea, sekulako galduko ez badira.

Baina, esan dugunez, hizkuntza-jardunari bagagozkio, ezin dugu ahaztu hainbat autorek nabarmendu dutena: hiztunaren baliabide linguistikoen errepertorioan, erabili ohi direnen errepertorioan, kolokazioak lokuzio edo esapide idiomatikoak baino gehiago dira, eta paremiak baino askoz ere gehiago. Horrek larriagotzen du aurreko lanetan nabari-nabaria den beste alderdi hau: kolokazioak dira, askogatik, gutxien bilduak eta landuak.

Horretaz jabetuta, Altzibarrek (2005) ahalegin handia egin du kolokazioen garrantzia nabarmentzeko. Corpasen ereduan lan eginez, kolokazioak

[17] http://www.hiru.com/hirupedia

kazetaritzaren diskurtsoan oso konbinazio erabiliak direla jarri du agerian, eta, horrekin batean, hainbat kolokazio-adibide bildu ditu, egitura sintaktikoaren arabera sailkatuta. Azken urteotan erabiltzen diren kolokazioak direla eta, badira, haren iritziz, hainbat arazo. Nagusia, honako hau (Altzibar, 2005: 12-13):

> ≪Kontua da, kolokazio(gai) asko eta asko berri samarrak izan arren, erabiliaren erabiliz indar hartzen, finkatze eta egonkortze bidean aurrera egiten ari direla, nahiz tradizioan sustrai gutxi edo argala duten, eta inguruko erdaren eredu eta moldez sortuak edo gutxienez haien eragin handikoak diren.≫

Zentzu horretan, Sarasolaren (1997) eta Garateren (1998) iritzi eta kritikekin bat egiten du Altzibarrek. Lan horietan, erdaren interferentziaz sortutako kalkoen hainbat adibide ematen dituzte, eta euskarazko kolokazio "naturalak" edo "jatorrak" zein diren aipatu. Gainera, gehienak `izena+aditza` motakoak dira, eta aditza kalkatzea da salatzen dutena. Adibidez, *atentzioa deitu*, *bilera ospatu*, *sermoia bota* eta *trena galdu*; aditz horiek gabe, beste hauek dirateke euskaraz erabiltzen direnak, edo duela gutxi arte nagusi zirenak: *atentzioa eman*, *bilera egin*, *sermoia egin* eta *trena hutsegin*.

Kezka hori zabalduta dago euskararen ikertzaileen artean; horren agergarri, UPV/EHUko Euskararen Institutuaren Kalkoen Behatokia proiektua dugu[18], zeinetan kalko fraseologikoen atala baitago. 40 kasu landu dira kolokazioen atalean, eta 14 lokuzioenean. `izena+aditza` osaerako adibide batzuk: **ilea hartu* (*adarra jo*) eta **oxigenoa eman* (*arnasa eman*) lokuzioak; **agurra jaso* (*agur egin*), **irteera hartu* (*irten*) eta **porrota jaso* (porrot egin) kolokazioak.

Altzibarrek tradizioa ondo ezagutzea eta albait gehien hari eustea aldarrikatzen du, baina aldi berean aitortzen tradizioak laguntzen ez digunean mailegu edo kalkoaren bidea egokia dela. Interesgarria da, gure ikerkuntzari bete-betean baitoakio, egin duen galdera hau:

> ≪Euskal kazetariek eta idazleek beren laborategian moldatu eta etengabe erabili eta hein batez beren erabileremuan behintzat hedatu dituzten kolokazioak hizkuntzaren norman finkatuak ote daude? Edo zein bai eta zein ez?≫

Horretarako, kolokazioen corpus eta hiztegien beharra azpimarratu du, eta, datu horiek izanik, kolokazioak eta lokuzioak bultzatzeko edo hobesteko irizpideak lantzea hizkuntzalariek.

[18] http://www.ehu.es/ehg/kalkoak/

Ildo horretatik, baina kolokazioez haraindi joanez, Altzibarrek euskarazko fraseologismo-hiztegi baten beharra aldarrikatu du (Altzibar, 2012). Hiztegia egiteko metodologia proposatu du, eta *a-aberats* bitarteko lagin bat ere landu, egin nahi lukeenaren erakusgarri.

Ikerketa enpirikoan ere badira zenbait lan aipatu beharrak. Batetik, Gurrutxagak (2002) zenbait izenen `izena+izenondoa` kolokazioak erauzi ditu, elkartze-neurriak erabiliz, *Orotariko Euskal Hiztegia* egiteko erabilitako testu-corpusetik eta *XX. mendeko Corpus Estatistiko*tik[19], eta emaitzak zenbait hiztegitan argitaratutako informazioarekin konparatu, kolokazio-erauzketak hiztegigintzan izango lukeen ekarria ikusarazteko. Gurrutxagaren (2003) master-mailako proiektuan, UFen ezaugarriak eta taxonomia landu ondoren, UFek euskarazko hiztegietan duten tratamendua aztertu du. Ondorioetako bat da kolokazioak hiztegiratzeko irizpideak eta metodologia landu behar direla; ariketa gisa, Elhuyarren hiztegigintzako datu-basean UFak, eta horien artean, kolokazioak errepresentatzeko eskema kontzeptuala proposatu du.

Etxebarria eta Bilbaok (2012) espezializazio-testuetan erabiltzen diren `magnitude-izena+aditza` egiturako kolokazioak identifikatu eta sailkatu dituzte. Datu-iturritzat, Elhuyar eta Ixa taldeak eratutako *Zientzia eta Teknologiaren Corpusa*[20] erabili dute (Areta et al., 2007). Bost magnitude aukeratu dituzte (*energia, abiadura, tenperatura, angelu, portzentaje*), eta bi kolokazio-mota bereizi: lexikoak (*energia askatu*) eta ez-lexikoak (*-z elikatu, -ekin erlazionatu, -n ibili*). Azken mota hori dela eta, gogoan izan hizkuntzalari batzuen iritziz aditzaren azpikategoriari (edo argumentu-egiturari) dagokiola (ikus II.1.1.1 atala). Aditzen aspektu lexikoa ere aztertu dute. Magnitudearen araberako aditzen azterketa konparatiboa egin dute, jakiteko zein aditz konbinatzen den magnitude guztiekin, batzuekin edo bakar batekin.

Fraseologia elebidunaren arloan, Villarrek (2011) alemanetik euskaratutako testuetan unitate fraseologikoen itzulpena aztertzeko erabiltzen ari den metodologia eta corpusa deskribatu ditu.

II.4 Euskarazko `izena+aditza` osaerako UFak

Gure ikergai diren `izena+aditza` osaerako UFen egitura, motak eta ezaugarriak sakonago aztertuko ditugu atal honetan. Badakigu horien artean esapide idiomatikoak (lokuzioak) eta kolokazioak ditugula. II.2.2.2 atalean,

[19] `http://xxmendea.euskaltzaindia.net/Corpus/`
[20] `http://www.ztcorpusa.net`

euskarazko kolokazioen eredu sintaktikoez mintzatu gara, gaingiroki. Lokuzioak dira, ordea, euskaraz gehien ikertu direnak, eta, horien artean, *aditz-lokuzio* kategoria da gure ikergai diren konbinazioak bere barnean hartzen dituena.

Aditz-lokuzioak aditz bat buru duten sintagmez osatuak dira (Urizar, 2012: 117). Urizarrek inplizituki berdintzen ditu aditz-lokuzioak Zabalak (2004) aztertzen dituen predikatu konplexuekin. Horren arabera, predikatu konplexuak dira aditz batez eta izen, adjektibo edo postposizio-sintagma batez[21] eraturiko hitz-segidak, generikoki [SX+Adi.] eran adierazten direnak[22], eta izaera predikatiboa duen sarrera lexikal gisa jokatzen dutenak. Zabalak zenbait adibide ematen ditu, definizioa ilustratzeko: *lan egin, hanka egin, min eman, loak hartu, bideari lotu, kontuan hartu, beldur izan* eta *zain egon*.

Zabalak azterketa sakon bat egin du, gai honetan egin diren ikerketak bilduz eta haietan oinarrituz. Honela laburbilduko ditugu alderdi nagusiak:

- Predikatu konplexuak eta *odolustu* moduko aditz-elkarteak kategoria edo mota desberdinekoak direla erakutsi du, Azkarateri jarraituz (Azkarate, 1990: 300-303). Horretarako froga sintaktikoak eman ditu, hala nola osagaiek modifikatuak eta bananduak izateko duten gaitasun desberdina. Ondorioz, morfologiaren eta sintaxiaren arteko muga-kasuak direla kontsideratzen du.

- Multzo ugaria baina mugatua da, hau da, zerrenda itxia osatzen dute.

- Ez dute kategoria edo multzo homogeneoa osatzen, ez dituztelako ezaugarri berberak, edo neurri desberdinean agertzen dituztelako. Hona hemen nagusiak:

  - Predikatu askotan, aditzaren eduki semantikoa murriztua da, edo oso oinarrizkoa. Batzuk aditz arinak edo euskarri-aditzak dira (*negar egin*); beste batzuk, laguntzaileak (*beldur izan*) edo loturaaditzak (*zain egon*).
  - Batzuen osaera ohiz kanpokoa da: aditzaren argumentuak mugatzailea izatea da ohikoa (*ardoa edan*), eta ez, *lan egin* predikatuan bezala, mugatzailerik ez izatea.
  - Sintaktikoki zurrunak dira, baina hein desberdinean. Izenaren inkorporazio-mailarekin lotu izan da portaera-desberdintasuna[23].

[21] Urizarrek dio adberbio bat ere izan daitekeela.
[22] Zabalak gaztelaniazko originalean darabilen [SX+V] egituraren euskarazko baliokidea.
[23] Ikus, adibidez, Martínez (1996).

Batzuek determinatzaileak edo adjektiboak onartzen dituzte (*negar asko egin, lan handia egin*), eta beste batzuek ez (*hanka egin, amore eman*). Foku-topiko mugimenduetan ere, portaera desberdinak daude: *Zertan egiten duzu lan? / *Nork egin du hitz*?

– Konposizionaltasun semantiko maila desberdinak dituzte. Konparazio baterako, *lan egin* predikatuaren esanahia oso erlazionatuta dago *lan*en ohiko esanahiarekin; *hanka egin*en esanahi literala, berriz, urrun dago 'joan' adieratik.

Predikatu konplexuak sailkatzeko, aditzaren ezaugarriak erabili ditu irizpidetzat. II.7 irudian laburbildu ditugu proposatutako multzoen eta azpimultzoen ezaugarriak.

Urizarrek beste irizpide bat erabili du aditz-lokuzioak sailkatzeko. Aditzarekin batera doan sintagmari erreparatuz, II.8 taulako sailkapena proposatu du.

Bi sailkapenak, oro har, hedadura bertsukoak dira, hau da, barnean hartzen dituzten egitura- edo konbinazio-motak nahiko bat datoz. Badira, ordea, aipatzeko moduko alde batzuk:

- Badirudi Urizarren sailkapeneko A2.2. (objektuaren predikatzailea + *ukan*) eta A.3 (subjektuaren predikatzailea + aditza) kategorien multzoa Zabalaren 2. kategoriaren ([XS+LOT/LAG]) baliokidea dela. Hala ere, Urizarren A.3.2 azpikategoria oso zabal definituta dago, ez da aditz-zerrenda itxi bat zehazten, eta barnean hartzen ditu Zabalaren definizioan sartzen ez diren lokuzio batzuk, hala nola *ados etorri, hauts bihurtu* eta *aiduru gelditu*.

- Zabalarentzat *egin* da aditz arin bakarra; Urizarrek, berriz, beste aditz batzuk ere sartu ditu kategoria horretan, nahiz eta aitortzen duen ez direla *egin* bezain emankorrak: *eman, hartu, esan* eta *jo*. Zabala jakitun da *eman* eta *hartu* aditz arintzat jo ohi direla, eta, zenbait arrazoi landu ditu horri aurka egiteko (Zabala, 2004: 481-483). Batetik, aditzaren adiera ez bide da asko urruntzen oinarrizko adieratik; bigarren, aditzak bere argumentu-egitura atxikitzen du; azkenik, XS sintagma beti da IS edo DS bat, eta, gainera, PS bat denean, ez du aditzaren argumentua asetzen (ez du objektu zuzena ordezkatzen).

Aurreko bi sailkapenak eta II.2.2.2 atalean eman ditugun kolokazio-egituren sailkapenek eskura jartzen dizkigute gure ikergai diren `izena+aditza` konbinazioen esparrua zehazteko osagaiak. Zehazte-lan hori diseinu esperimentalaren barnean egingo dugu (ikus IV.2 atala).

1 [XS+Adi] multzoa: aditzak ematen dio konplexuari izaera predikatiboa, eta XS elementua predikatu horren argumentua edo modifikatzailea da. Horien artean, hiru azpimultzo bereizten ditu:

   1. 1 [XS+Adi$_{arina}$]: *lan egin, turrut egin, itxurak egin, haginka egin*
   1. 2 [IS/DS+Adi]: *min hartu, hitz eman, hanka sartu, hitza hautsi, begiak zorroztu, kontuak atera, txaloak jo*
   1. 3 [S-kasua+Adi], [PS+Adi]: *buruak eman, goseak egon, bideari lotu, negarrari eman; hegaz egin, harira etorri, larrutik pagatu, boladan egon, kontuan hartu, aintzakotzat hartu*

2 [XS+LOT/LAG] multzoa: XS elementuak du predikatu gisa behar den informazio guztia, hau da, izen-predikatua da, eta berak ematen dio argumentu-egitura konplexuari; aditza, berriz, beharrezkoa da predikatuak bere argumentu-egitura sintaxian proiekta dezan, hau da, lotura-aditz edo laguntzaile hutsa da. XS elementuaren kategoriaren eta aditzaren arabera osatzen ditu azpimultzoak:

   2. 1 [IS+*izan*]: *beldur izan*
   2. 2 [IS+ °*edun*]: *damu °edun*
   2. 3 [AdjS+*izan*]: *posible izan*
   2. 4 [AdjS+°*edun*]: *maite °edun*
   2. 5 [IS+*egon*]: *giro egon, zain egon*
   2. 6 [AdjS+*egon*]: *oker egon*
   2. 7 [AdjS+*ibili*]: *eske ibili, oker ibili*

II.7 Irudia: Zabalaren (2004) predikatu konplexuen sailkapena.

Bada, ordea, hemen aztertzekoa den gogoetagai bat, Urizarren aditz-lokuzioen eta Zabalaren predikatu konplexuen idiomatikotasunarekin eta, beraz, UFen sailkapenarekin zerikusi zuzena duena. Lehen begiratu batean, badirudi ez dugula zertan kuestionatu horiek denak esapide idiomatikotzat jotzea. Hala ere, badira sakonago aztertzea merezi duten alderdi batzuk.

Aurrerago esan dugunez, Zabalak argi adierazi du predikatu konplexu

A `Izen- edo adjektibo-sintagma biluzia + aditza`

- A.1 Objektu zuzena + aditza
  - A.1.1 ***min egin*** (XS + aditz arina)
  - A.1.2 ***beldur dio*** (XS + *ukan*)
- A.2 Objektuaren predikatzailea + aditza
  - A.2.1 ***atsegin du*** (XS + *ukan*)
  - A.2.2 ***gogait egin*** (XS + aditz arina)
- A.3 Subjektuaren predikatzailea + aditza
  - A.3.1 ***falta da*** (XS + *izan/egon*)
  - A.3.2 ***ados etorri*** (XS + bestelako aditza)

B `Izen- edo adjektibo-sintagma mugatua + aditza`

- B.1 Nominatibodun sintagma + aditza — ***adarra jo***
- B.2 Ergatibodun sintagma + aditza — ***suak hartu***
- B.3 Datibodun sintagma + aditza — ***hitzari eutsi***

C `Adizlaguna edo aditzondoa + aditza`

- C.1 Argumentua + *egin* — ***hegaz egin, haginka egin***
- C.2 PS + bestelako aditza — ***aurrera eraman, gogora etorri***

II.8 Irudia: Aditz-lokuzioen sailkapena (Urizar, 2012: 119).

guztiak ez direla konposizionaltasun semantiko maila berekoak, eta zurruntasun sintaktikoa ere aldatzen dela batetik bestera. Horretan bat dator Rodriguez eta Murgaren (2003) iritziarekin, honela baitiote, `izena+egin-arin` predikatuez dihardutela:

> ≪Badira adierazpen konplexu batzuk beste batzuk baino modifikazio sintaktiko handiagoa onartzen dutenak. Era berean, badira egitura batzuk besteak baino opakoagoak direnak, esanahiari dagokionean.≫

Beraz, gure terminologia erabiliz, predikatu konplexuek ez dute idiomatikotasun-maila bera. Orduan, esapide idiomatikoak eta kolokazioak nola

definitu ditugun kontuan izanik, lehen galdera bat hau da: euskarazko ikerlanetan estandartzat jo daitekeen "lokuzio" terminoaz etiketatu ohi diren guztiak esapide idiomatikoen kategorian sailkatzekoak dira, ala badira batzuk kolokazioen kategorian egokiago leudekeenak?

Batetik, Zabalaren errepertorioan ageri diren batzuek zalantza sor dezakete; hauek, esate baterako: *arnasa hartu, atentzioa eman, hasiera eman, eredutzat hartu, euria egin.* Horietan, kolokazioetan bezala, oinarriak bere ohiko edo oinarrizko adiera atxikitzen du, eta nolabaiteko malgutasun sintaktikoa ere badute, bakoitzak bere neurrian. Adibidez:

(1) Ez da gutxi, txuri-urdinek gainean zuten presio neurrigabea kontuan hartzen bada. Zinez estimatuko dute atzo **hartutako arnasa**[24].

(2) Adibidez, niri **atentzio handia eman** dit Espainia izateak arto transgeniko gehien egiten duen herrialdea Europan[25].

(3) Hots, eskolak aldatu egin behar du, ezin du hain transmisiokoa izan, ezin du **eredu bakartzat** testu-liburua **hartu**, ezin du hain kalifikatzaile ona izan[26].

(4) Ekaineko lehen hamabostaldian **egindako euriek** bete-betean harrapatu zuten mahastien loraldia[27].

Altzibarren (2005) kolokazio-adibideen artean ere, *atentzioa eman* eta *hasiera eman* daude.

Horiek horrela, esango genuke horrelakoek gehiago dutela kolokaziotik esapide idiomatikotik baino (semantikoki zein sintaktikoki "erdiidiomatikoak" dira). Gutxienez, esapide idiomatikoen eta kolokazioen faseartean leudeke, baina, inon sailkatzekotan, uste dugu egokiagoa dela kolokaziotzat hartzea.

Bestetik, Urizarren A multzoko unitateak ditugu, batzuk (*lan egin, min hartu*) Zabalaren [XS+Adi] multzoan daudenak; eta beste batzuk (*falta izan*) [XS+LOT/LAG] multzoan. Ezaugarri batzuetan, horien profila ez da II.2 atalean definitu ditugun ohiko esapide idiomatikoen modukoa. Maiztasun handikoak dira eta, gehienez, erdikonposizionalak. Contreras eta Suñerrek ez dituzte *lan egin*, *hitz egin* eta kidekoak idiomatikotzat jotzen (Contreras eta Suñer, 2004: 86-87). Idiosinkrasia sintaktikoa eta finkapena da horien

[24] http://paperekoa.berria.info/kirola/2005-12-22/026/857/.htm

[25] http://www.zientzia.net/informazioa/elhuyar/2009/253/Pdf/30-45.pdf

[26] Eskola porrota: Akademizismotik ihes egiteko garaia (Aldizkari-atala). Komunikazio Biziagoa, S.A.L, 2009. Egileak: GARCIA, Mikel. Aldizkaria: Argia, 2009-09-27.

[27] Etorkizuneko sagardoa, Eusko Labelarekin (Aldizkari-atala). Euskal Irrati Telebista, 2011. Egileak: GALARRETA, Xabier. Aldizkaria: Sustraia, 2011-01-01.

bereizgarria, denek maila berekoa ez badute ere. Interesgarria da hemen gogoraraztea Tutinen iritziz frantsesezko kolokazio batzuk era atipikoan edo arkaikoan eratuak direla (*avoir faim, avoir soif, rendre visite, avoir rendez-vous*), ez baita ohikoa, hizkuntzaren arau estandarren arabera, horrelakoetan izenak determinatzailerik ez izatea (Tutin eta Grossmann, 2002: 8-9).

Beraz, horrelakoak sailkatzeko irizpide desberdinek eragina izan dezakete karakterizazioan, eta, interesgarria izan daiteke aztertzea zer eragin duen horrek emaitzetan. Printzipioz, horrelakoak sistematikoki esapide idiomatikotzat sailkatzea baino egokiagoa izan daiteke kasuan-kasuan erabakitzea, beste unitateetarako ezarriko diren sailkatze-irizpide orokorren arabera (VI.4 atala).

Dena den, aditz arinen kasuan, komeni da zehaztapen batzuk hemen egitea, horrelako aditzak dituzten UFen ezaugarri semantikoen eta morfosintaktikoen inguruan.

Aditz arineko UFak

Aditz arinak edo euskarri-aditzak gaingiroki aurkeztu ditugu kolokazioen murrizketa lexikalaz eta konposizionaltasunaz jardun dugun II.2.2.3 atalean. Eman ohi den definizio estandarraren arabera, aditz arinak tematikoki osatugabeak dira (Grimshaw eta Mester, 1988: 205), eta osagarri bat behar dute subjektuari rol tematiko bat esleitzeko (Zabala, 2004: 455), edo aspektu-informazioa gehitzeko (Rafel, 2004: 396). Alonsok (1998) adibide hauek ematen ditu: *to take a walk, dar un paseo, faire une promenade* (euskaraz, *paseoa eman, buelta bat eman, ostera bat egin* izan daitezke ordainetako batzuk). Eraikuntza horietan[28], aditz arinak eduki semantiko gutxi edo murriztua du; beraz, osagarriak ematen dio balio predikatiboa eraikuntzari, eta osagarritik aditzerako argumentu-transferentzia gertatzen da (Grimshaw eta Mester, 1988: 211). Eskuarki, osagarria izena dela esan ohi da, baina badira adjektiboa edo postposizio-sintagma ere izan daitekeela diotenak (Contreras eta Suñer, 2004: 87; Wotjak, 1998: 267; Krenn, 2008).

Horrelako eraikuntzetan agertzen diren aditzak iragankorrak izaten dira, eta multzo mugatua osatzen dute (Wotjak, 1998: 267). Maiztasun handiko aditzak dira, eta esanahi generiko edo oinarrizkoa izaten dute; adibidez,

[28]Horrelako eraikuntza edo konbinazioak izendatzeko, termino bat baino gehiago erabili ohi dira; bestak beste, *light verb construction* ("aditz arineko eraikuntza"), *support verb construction* ("euskarri-aditzeko eraikuntza"), *construcción verbo-nominal funcional* (Wotjak, 1998), *verbo compuesto* (Koike, 1998: 254), *semi-compositional construction* (Vincze, 2013), *construcción con verbo de apoyo*, *predicado complejo* (Alonso, 2000), *Funktionsverbe* (Von Polenz, 1963).

ingelesez: *do, give, have, make, take*; gaztelaniaz, *dar, hacer, poner, coger, echar*; frantsesez: *avoir, faire, donner, prendre*. Urizarrek euskarazko hauek aipatzen ditu: *egin, eman, hartu, jo, kendu, ukan*.

Aditz arineko eraikuntza askori osagarriarekin morfologikoki erlazionatua dagoen aditz bakun bat dagokie (Alonso, 1998). Euskaraz ere aipatua da fenomeno hau (Etxepare, 2003: 27). Adibidez, *mehatxu egin / mehatxatu*; *dantza egin / dantzatu*; *agur esan / agurtu*. Maiz aditz bakuna predikatu konplexuaren baliokidea edo esanahi bertsukoa dela esaten bada ere, oso gutxitan dira erabat baliokide edo sinonimoak, era batzuetako desberdintasunak izaten baitira (Urizar, 2012: 120)

Orain arte esandakoa kontuan izanik, badirudi aditz arineko eraikuntzetan izenak bere ohiko edo oinarrizko esanahia atxikitzen duela. Izan ere, hala ez balitz, osagarriak aditzari transferitzen dion balio predikatiboa konbinazioaren balio figuratibotik letorke, hau da, konbinazioan soilik duen balio batetik, sorgin-gurpil bat osatuz. Gure ikuspegitik, *kale egin* edo *hanka egin* lokuzioetan, ez da hain argi zein den *kale* eta *hanka* izenei legokiekeen egoera, ekintza edo balio predikatiboa, eta zer argumentu-egitura transferi liezaiokeen lokuzioari. Hor *egin* aditzak aditz arin baten antzeko funtzioa izan arren, lokuzioaren izaera eta esanahia ezin da transferentzia-mekanismo horren bidez esplikatu. Hala aitortu du Zabalak ere, *kale egin*en kasuan forma eta esanahiaren erlazioa erabat arbitrarioa dela dioenean (Zabala, 2004: 469), edo *bihotz eman* ('animatu') zein *gogo hartu*ren ('erabaki') esanahiak osagaien esanahietatik ondorioztatzea erraza ez dela dioenean (Zabala, 2004: 486).

Ondorioz, aditz arinen eta haien eraikuntzen definizioaren gunean mekanismo hori sartu dugunez, horrelako eraikuntzak semantikoki erdikonposizionalak lirateke. Beste ezaugarri aski onartu bat da horrelako eraikuntzek malgutasun morfosintaktiko erlatiboki handia izaten dutela (Wotjak, 1998: 267; Vincze, 2013: 6-7). Euskaraz ere hori gertatzen dela azaldua dugu dagoeneko. Horiek horrela, badirudi, kolokazioekin batera, esapide erdiidiomatikoen kategoriakoak liratekeela, ez esapide idiomatikoen kategoriakoak.

Beste puntu garrantzitsu bat da horien malgutasun morfosintaktikoa neurtzerakoan zer onartuko den aldakuntzatzat. Hain zuzen ere, Urizarrek dio lokuzio askok, hizkuntza-sistemaren eragiketak aplikatuz gero, lokuzio-izaera galtzen dutela, eta unitatearen esanahia ere alda daitekeela (Urizar, 2012: 105). Horretara, *lan bat egin dut* edo *egin dudan lana* esaldiak ez legozkioke *lan egin* aditz-lokuzioari, bi arrazoirengatik: a) *lan* hitza zenbakaitz izatetik zenbakarrira izatera igaro da; eta b) lokuzio-izaerak ematen dion kohesioa galtzen du, eta kolokazio bihurtzen. Arrazoi beragatik lirateke *atsedena hartu dut* eta *atseden ederra hartu dut atsedena hartu* kolokazioa-

ren agerpenak, eta ez *atseden hartu* lokuzioarenak. Aurrekoen antzera, bi esaldi hauek ere egitura desberdinekoak lirateke, (5a) adibideko *egin* ez baita aditz arintzat jotzen, "astuntzat" baizik (Rodriguez eta Murga, 2003: 417-418; Zabala, 2004: 467-468).

(5) a. Jonek lana egin du
b. Jonek lan egin du

Gure iritziz, planteamendu horrek arazo eta kontraesan batzuk dakartza berekin.

Lehenik, ondorengo erabilera hau ez dator bat bereizketa-eskema horrekin: *zuk non egiten duzu lana?* Izan ere, semantikoki (5b)ko *lan egin* bezalakoa da, baina morfosintaktikoki, (5a) bezalakoa. *Lana egin* konbinazioaren hainbat adibide daude *Orotariko Euskal Hiztegi*an (OEH), *lan egin* azpisarreraren barnean, hau da, horren aldaeratzat emanak, eta nabarmentzekoa da, gainera, semantikoki ere (5b)ko *lan egin* bezalakoak direla:

(6) Lana egiten duenak saria edo jornala merezi du. Lard 393 .

(7) Lana egin bearrian, denbora alperrik galtzen (V-gip, G-azp, AN-gip) Gte Erd 157.

Horren arabera, *lan/lana egin* aldaerak ditugu esku artean. Horrelako gehiago ere badaude; esaterako, *behar/beharra egin* eta *eztul/eztula egin*. *Behar*en kasuan, OEHk dio *beharra egin* ugariagoa dela. *Eztul*en kasuan, bi aldaerak baliokideak dira semantikoki, *eztul* ez baita izen zenbakarritzat erabiltzen: **bi eztul egin* $\longrightarrow$ *bi aldiz eztul(a) egin*.

Bigarren, ikusi dugu, Zabalaren lana azaltzean, euskarazko predikatu konplexuek zurruntasun morfosintaktiko desberdina izaten dutela, eta izenaren inkorporazio-mailarekin erlazionatu dela hori. Batik bat, `izena+egin` predikatuak aztertu dira, eta oso eztabaidatua da horien estatusa.

Bada, inkorporazioaren alde (Uribe-Etxebarria, 1989; Oyharçabal, 1994; Martínez, 1996; Fernandez, 1997) eta kontra (Levin, 1983; Ortiz de Urbina, 1989; Laka, 1993) eman diren argudioen artean, horrelako aldakuntzen ebidentziak erabili dira (*lan ederra egin, lo gutxi egin*). Rodriguez eta Murgaren lanean bertan, *Mikelek lan handia egin du* adibidea ematen da, *lan egin* egiturak izenondoaren modifikazioa onartzen duela erakusteko (Rodriguez eta Murga, 2003: 421).

Oyharçabalen (2006) ondorioa da hiztun askorentzat (8)ko hiru egiturak erabilgarri daudela, baina, Martínezen (1996) ildo beretik, erabilgarritasun hori predikatu konplexuaren arabera eta euskalkiaren arabera aldatzen dela.

(8) a. Lan ederra/gutxi egin dugu (DP + absolutive)
b. Ederki/gutxi egin dugu lan (NP + inherent case)
c. Ederki/gutxi lan egin dugu (N_{inc})

Baina horiek *lan egin*, *lo egin* eta abarren agerpenak ez balira, ez luke zentzurik haien egitura sintaktikoaren analisian eta inkorporazio-mailaren eztabaidan kontuan hartzeak edo argudiotzat erabiltzeak ere, ez bailirateke esanguratsuak.

Sintaxiaren arloan egindako ikerketa horiek aintzat hartuz, zentzuzkoa dirudi kontsideratzeak aldakuntza horiek esanguratsuak direla predikatu konplexu bakoitzaren izaera ebazterakoan. Hori onartzera, ezin esan genezake *lan ederra egin du* agerpena *lan egin du* unitatearen aldakuntza ez denik, eta kontuan hartu behar genuke unitatea sailkatzerakoan zein haren idiomatikotasun morfosintaktikoa neurtzean.

II.5 Laburpena

Kapitulu honetan, idiomatikotasunaren kontzeptuaren zenbait interpretazio aztertu ondoren, gure ikerkuntzarako definizio operatiboa zehaztu dugu, zeinen arabera idiomatikotasuna UFen ezaugarri definitzailea baita, konplexu eta graduala, eta lau osagai edo propietate hauek osatua: instituzionalizazioa, idiosinkrasia estatistikoarekin lotua; ez-konposizionaltasun semantikoa, edo konposizionaltasun partziala; eta finkapen morfosintaktiko eta lexikala. Erabaki horrek behartuko gaitu propietate horiek neurtzeko metodologietan sakontzera, eta neurketa horien emaitza modu integratuan erabiltzera, idiomatikotasuna karakterizatuko badugu.

Lau propietate horien konbinazioak sortzen duen idiomatikotasun-continuumean, bi zona edo eremu zabal daudela erakutsi dugu: esapide idiomatikoak eta kolokazioak (edo erdiidiomatikoak). Horietako bakoitzean ere, autore batzuek azpikategoriak bereizi dituzte; batetik, esapide idiomatiko opakoak eta figuratiboak; bestetik, kolokazio murriztuak eta irekiak. Bi azpikategoria-bikote horien barneko bereizketa aski lausoa dela aitortua dago, eta horrek eragina izan dezake sailkapen-ereduaren aplikagarritasunean.

Azkenik, euskarazko fraseologiaren ikuspegi arinaren ondoren, ikergai ditugun `izena+aditza` osaerako UFen ezaugarriak aztertu ditugu, eta, bereziki, aditz arineko eraikuntzak jorratu ditugu. Horiek sailkatzeko irizpideak malgutu beharra eta malgutasun morfosintaktikoa neurtzeko aldakuntzak kontuan hartu beharra dira atera ditugun diseinu esperimentalari begirako ondorio nagusiak.

III. KAPITULUA

UFen erauzketa eta karakterizazio automatikoa

Kapitulu honetan, ikergaitzat hartu dugun egitekoa, UFen erauzketa eta karakterizazio automatikoa, deskribatuko dugu, haren helburuak, urratsak eta ebaluazio-metodologiak azalduz, eta idiomatikotasunaren propietate bakoitza neurtzeko zein neurketa horiek konbinazioan erabiltzeko garatu diren tekniken berri emanez.

III.1 UFen erauzketa, fraseologia konputazionalaren egitekoetako bat

Fraseologia konputazionala unitate fraseologikoen prozesamendu automatikoaz arduratzen den hizkuntzalaritza konputazionalaren atala dela esan liteke. Prozesamendu horren barnean, hainbat egiteko edo ataza deskribatu dira. Heiden (2008)[1], eta Baldwin eta Kimen (2010) proposamenetan oinarrituta, Urizarrek honako ataza hauek bereizten ditu (Urizar, 2012: 169-180):

- **Erauzketa**. Testuetatik UFak eskuratzea, eskuarki lexikoiak edo hiztegiak elikatzeko.
- **Identifikazioa**. UFen testuetako banakako agerpenak atzematea. Identifikazioaren egitekoa da konbinazio baten UF erako interpretazioa eta hitzez hitzeko interpretazioa duten agerpenak bereiztea.

[1] Fraseologia konputazionala terminoa (zehazki, *Computational Phraseology*), Heiden (2008) lanean aurkitu dugu lehen aldiz literaturan. Jakina, kontzeptua ez da Heidek sortua, baina bai, ordea, bataiatua eta lehen aldiz sistematikoki landua.

- **Interpretazioa**. UFen barne-sintaxia eta semantika desanbiguatzea. Erauzketaren eta identifikazioaren parte gisa aurkezten du.
- **Deskribapena**. UFen malgutasun morfosintaktikoaren berri zehatza ematea. Oso lotuta dago identifikazioaren atazarekin.

Tesi-lan honen helburua euskarazko UFen erauzketa eta karakterizazioa ikertzea denez, lehen egitekoaren arloan egindako ikerkuntza- eta garapenlanak xehe azalduko ditugu kapitulu honetan.

III.2 UFen erauzketaren helburuak eta urratsak

Adierazi berri dugunez, erauzketaren helburu nagusia da testuetatik UFak eskuratzea. UFaren definizioa egin dugunean, esan dugu nolabaiteko "unitate aurrefabrikatuak" direla, konposizionaltasun-printzipioaren arabera soilik ezin aurreikusi edo azaldu daitezkeenak. Pentsa genezake unitate horiek aski finkatuak direla, ez direla hizkuntza-jardueran konposizionalki osatzen ditugun konbinazio libreen modukoak, eta, deskribapen-maila minimo bat erdietsi duen hizkuntza "heldu" baten kasuan, dagoeneko bilduak eta deskribatuak egongo direla hiztegi eta antzeko baliabideetan. Orduan, galdetu egin behar genuke zergatik ustiatu behar diren testuak UFak erauzteko. UF berriak lortuko ditugu? Informazio gehiago edo aberatsagoa lortuko dugu lehendik bilduak ditugun UFei buruz? Aurreko bi galderen erantzunak baiezkoak dira. Ikus dezagun zergatik diogun hori.

- UFek ez dute denboran aldatzen ez den multzo itxi bat osatzen, hizkuntzaren erabileraren azterketak UF berriak azaleratzen ditu; fenomeno hau bereziki nabaria da kolokazioetan, eta, hein txikiagoan, esapide idiomatiko figuratiboetan (esapide idiomatiko opakoen sailean, berriz, nobedade gutxiago izaten da).

  Erabilerak dakartza berritasun horiek, eta horien ebidentzia corpusetan aurki dezakegu. Gogoan izan behar dugu UFen, eta bereziki kolokazioen, funtsezko ezaugarrietako bat idiosinkrasia estatistikoa dela, eta, hori neurtu ahal izateko, ikertzaileak datuak behar dituela, hizkuntza-lagin kuantitatiboki adierazgarriak, hain zuzen ere. Horretarako, hizkuntzalaritzaren beste ikerketa-ildo batzuetan erabiltzen den introspekzioa ez da aski, ez aski fidagarria eta adierazgarria behintzat.

- Erauzketaren bidez informazio berria lor dezakegu lehendik bilduak ditugun UFei buruz:

- Lehen datu garrantzitsua da UF bakoitzak erabilera errealean duen benetako pisua; erabiltzen den edo ez, eta zenbateraino.
- Erauzketan sailkapena integratzen dugunean, UF bakoitza zein mota edo kategoriatakoa den ere jakin dezakegu: esapide idiomatikoa, kolokazioa... Karakerizazio hori garrantzitsua da UFak hiztegietan desberdin antolatu ohi direlako, edo, itzulpen automatikoari begira, prozesamendu desberdinak komeni izan daitezkeelako.
- Aurrerago ikusiko dugunez, erauzketa-teknika batzuek konbinazioen malgutasun morfosintaktikoa eta lexikala neurtuz ekiten diote idiomatikotasunaren karakterizazioari, eta horretarako darabilten informazioa baliagarria izan daiteke UFen deskribapen zehatza egiteko; adibidez, hizkuntzaren prozesamendu automatikoan beharrezkoak diren datu-base lexikaletan.

Beraz, ez da harritzekoa corpusetatik UFak automatikoki eskuratzeko teknikak garatu izana.

Erauzketaren helburuak azalduta, eskuratze-prozesuaren urratsak aztertuko ditugu. Bi urrats bereizi ohi dira (Seretan, 2011: 31):

- UFak hitz-konbinazioak dira, ezaugarri jakin batzuk dituzten konbinazioak izan ere, eta, corpus batean dauden UFak eskuratuko baditugu, begien bistakoa da hau dela lehen urratsa: UF izan litezkeen hitz-konbinazioak ateratzea, UF izateko konbinazio hautagaiak, hain zuzen ere. Urrats horri *UF hautagaien erauzketa* deritzo.

- Hautagai horietako asko ez dira unitate fraseologiko, konbinazio libreak baizik. Bestetik, II.2 atalean azaldu dugunez, UF guztiak ez dira idiomatikotasun-maila berekoak, eta mota batzuk bereizi ditugu, nagusiak esapide idiomatikoak eta kolokazioak izanik. Bada, eskuratze-prozesuaren bigarren urratsa da erauzitako hautagaiak bere idiomatikotasun-mailaren arabera automatikoki antolatzea (rankingetan, esaterako) edo sailkatzea. Horri esaten diogu *UFen idiomatikotasuna automatikoki karakterizatzea*.

Bi urrats horiek ez dira beti erabat argi bereizten. Esaterako, hautagaien erauzketan, informazio estatistikoa erabiltzen da sistema batzuetan, atari bat gainditzen ez duten hautagaiak bazter utziz. Bestetik, sistema batzuek (Smadja, 1993) bigarren urratsaren ondoren erabiltzen dute hautagaiak erauztean erabili ohi den informazio morfosintaktikoa, lehen ho-

rrelakorik ezarri gabe neurri estatistikoekin sortutako hautagai-rankingak iragazteko.

III.3 UF hautagaien erauzketa

UF hautagaiak biltzeko oinarrizko prozedura da hitz baten eta haren inguruan agertzen diren hitzen konbinazioak eratzea. Hori egin ahala, gainera, konbinazio bakoitza zenbat aldiz agertzen den zenbatzen da (agerkidetza-maiztasuna).

Konbinazioak sortzeko era zehatza, ordea, zenbait irizpide edo parametroren araberakoa izaten da:

- Konbinazioaren osagaien izaera: lemak, hitz-formak, bestelako informazio morfosintaktikoa (kasua, mugatasuna...).
- Hitzen arteko konbinazioak osatzeko bete behar diren baldintzak. Bi irizpide ezar daitezke:
  - Distantzia: hitz bat beste batetik zenbateko distantzia maximora ager daitekeen, harekiko agerkidetza edo konbinaziotzat jotzeko. Hitz baten ezker-eskuin "leiho" (*window*) bat eratzen da, eta horren zabalera (*span*) zehazten (halako token-kopurua).
  - Konbinazioaren osagaien arteko erlazio morfosintaktikoa. Distantzia gorabehera, agerkidetza bat egiturazko erlazio baten instantzia bat da (Evert, 2008: 19).

Prozesamendu linguistikorik gabeko testu gordina (*raw text*) erabiltzen denean, hitzen formen konbinazioak soilik egin daitezke, eta, gainera, ez dago modurik osagaien arteko erlazio morfosintaktikoaren irizpidea erabiltzeko. Distantzia da konbinazioak osatzeko modula dezakegun parametro bakarra. Horrelako oinarrizko sistemek hainbat arazo dituzte.

Lehenik, flexio-sistema minimo bat duten hizkuntzetan, eta, bereziki, hizkuntza eranskarietan, azaleko formen konbinazioak eratzeak *sparse data problem* edo "datu-barreiatzearen arazoa" dakar berekin (Heid, 2008). Esana dugu UFak ez direla hitzen formen konbinazio erabat finkoak. UF batzuk, batez ere kolokazioek, baina esapide idiomatiko batzuek ere bai, aldakuntza morfosintaktikoak izan ditzakete; horietatik sinpleenetakoa, osagaien flexioa. Ondorioz, lema beraren agerpenak zenbait formatan banatuta daude, hau da, datuak barreiaturik daude, eta horrek kasu askotan urritasunera darama, batez ere maiztasun handikoak ez diren hitzen kasuan.

Hainbat autorek erakutsi dutenez (Heid, 1994: 249; Stubbs, 2001: 82-83; Evert, 2008: 35), forma flexionatuen agerpenak dagokien lemarekin multzokaturik prozesatzeak emaitza estatistiko esanguratsuagoak lortzeko aukera ematen du. Tesi-lan honetan ikergaitzat dugun `izena+aditza` osaerako UFetan, flexioa nahitaez prozesatu beharra dago, baldin azaleko forma desberdinak forma kanoniko bakarrarekin erlazionatzeko gai izango bagara. Adibide batzuk:

(9) **Adarra jotzeko** gogoz, ala? Ez **adarrik jo** niri.

(10) Atzoko partidan hiru **gol sartu** zituen. **Gola sartzen** ahalegindu da. Horrela jarraituz gero, ez du **golik sartuko**. Nork **sartu** ditu gaurko bi **golak**?

Lehen adibide-parea *adarra jo* esapide idiomatikoarena da, eta bigarren saila, berriz, *gola sartu* kolokazioarena.

Baina UFen forma kanonikoan lema-konbinazio hutsak erabiltzeak ere baditu bere desabantailak. Ohartu behar gara forma kanonikoa ez dela lemen konbinazio hutsa. UFaren forma kanonikoan, haren egitura sintaktikoaren arabera, osagai batzuen lema agertuko da, eta beste batzuen forma. Aurreko adibideetan, lema-konbinazioak *adar jo* eta *gol sartu* dira, baina UFarentzat proposatu ditugun forma kanonikoak bestelakoak dira (*adarra jo*, *gola sartu*). Are argiago ikusiko dugu arazo hau euskarazko *adarretatik heldu*, *gogora ekarri* eta *arriskuan jarri* UFei erreparatuta. Horien lema-konbinazioetan leudekeen lemak *adar*, *gogo* eta *arrisku* dira, eta bistan da forma kanonikoak horietatik sortuko bagenitu, oso informazio garrantzitsua galduko genukeela. Are gehiago, *kontu izan* eta *kontuan izan* UFei lema-konbinazio bera dagokie (*kontu izan*), eta ez dira UF bera. Bistan da, beraz, `izena+aditza` osaerako UF baten forma kanonikoan, kasua definitzailea dela, eta mugatasuna ere kontuan hartzekoa dela.

Ondorioz, esan genezake konbinazioaren osagaien flexioak eragindako aldakuntzak lemaren informazioaren bidez prozesatzea komeni dela, eta, forma kanonikoa automatikoki erauzteko, kasua eta mugatasuna ere kontsideratu behar direla.

Bigarren, UFak ez dira izaten edozein kategoria-konbinaziotakoak, egitura sintaktiko batzuen araberakoak baizik, eta hori oso nabaria da kolokazioen kasuan, II.2.2.2 atalean Altzibar eta Urizarren lana azaltzean adierazi dugunez (Altzibar, 2005: 4-12; Urizar, 2012: 99). Kolokazioak honelako egituretako bat izan dezaketen konbinazio bitartzat definitu izan dira: `izena+adjektiboa`, `izena+izena`, `izena+aditza`, `adberbioa+adjektiboa`, `adberbioa+aditza`... Beraz, alderdi hori kontuan hartu ezean, testu hu-

tsetik lortuko ditugun konbinazio estatistiko "esanguratsu" asko ez dira linguistikoki errelebanteak izango, eta UF hautagai izateko aukera gutxi izango dute.

Horrelako emaitza ez-interesgarri batzuk *stop-word* direlakoak iragaziz saihets daitezke, baina horrek ez du arazoa era sendoan konpontzen (eta are gutxiago lehen flexioaz jardutean azaldu dugun datu-barreiatzearena), ez baitigu aukerarik ematen egitura morfosintaktiko jakin bateko UFak erauzteko. Horretarako, gutxienez, osagaien kategoria morfosintaktikoaren informazioa behar dugu. Informazio hori etiketatze morfosintaktiko (*POS tagging*) izeneko prozesuaren bidez lortzen da.

Horrenbestez, aski onartua da UFak erauzteko sistema baten abiaburu-baldintza dela testuak prozesamendu linguistiko minimo bat izatea (Seretan, 2011: 44). Minimo hori hizkuntzaren arabera aldatzen da, baina, oro har, esan daiteke lematizazioa eta etiketatze morfosintaktikoa direla minimo hori adierazten duten estandarrak. Horiek osatzen dute sistema gehienetan erabiltzen den prozesamendu linguistikoa (Heid, 2008: 351).

Informazio hori edukita, aukera sinple bat da kategoria jakin batzuetako hitzen arteko konbinazioak osatzea. Esaterako, `izena+aditza`, `izena+izena`, `izenlaguna+izena+izenondoa`, eta abar. Dena den, kategoria-segida hutsa baldintza arinegia izan daiteke hitzen arteko erlazio sintaktikoa dagoela ziurtatzeko, batez ere hizkuntza batzuetan. Esaterako, *oparitu zenidan liburua espero nuen baino gehiago gustatu zitzaidan* esaldian, *liburua* izena eta *espero izan* aditza segidan gertatzeak ez du adierazi nahi erlazio sintaktikoa dutenik eta konbinazio "erreal" bat osatzen dutenik, *liburu espero izan* agerkidetza hutsa baino ez baita.

Horregatik, azken urteetan, zenbait ikertzailek aldarrikatu dute UFen erauzketa-sistema baten eraginkortasuna analisi sintaktikoaren bidez hobetu daitekeela (Goldman et al., 2001; Pearce, 2002; Pecina, 2009), hau da, lehen aipatu dugun erlazio morfosintaktikoaren alderdia etiketatze morfosintaktiko hutsetik harantz eramatea, kategoria-segida jakin batzuen araberako konbinazioak osatzetik, azaleko (*shallow parsing*) zein sakoneko (*deep parsing*) analisi sintaktikoaren arabera halako erlazio sintaktikoa duten osagaien arteko konbinazioak osatzera. Beraz, azaleko murriztapen-gramatika baten bidez edo baterakuntza-gramatika baten bidez, erlazio morfosintaktiko jakin bat duten hitzak erabiliz sortzen dira konbinazio hautagaiak.

Seretan da analisi sintaktikoak kolokazio-erauzketan duen eginkizun giltzarria gehien nabarmendu duen autoreetako bat (Seretan eta Wehrli, 2006; Seretan, 2011). Fips izeneko analizatzaile sintaktiko sakona erabiltzen du (Wehrli, 2007), `izena+aditza` osaerako kolokazioen erauzketan, eta hobekuntza nabaria dela aldarrikatzen du, emaitzak leihoaren sistema eta kate-

goria gramatikalen segidaren araberako iragazketa konbinatuz lortzen den oinarri-lerroarekin konparatuz.

Hala ere, berak aitortu du (Seretan, 2011: 47) azaleko analisi sintaktikoa edo horren eta sakonaren artekoa den mendekotasun-analisia erabiliagoak direla analisi sakona baino. Arrazoi nagusia bide da analizatzaile sakonen fidagarritasun eta eskuragarritasuna txikiagoa dela, batez ere hizkuntza batzuetan. Deigarritzat du, dena den, ingelesean horrelako analizatzaileak ugariak eta aski fidagarriak izan arren, hizkuntza horretako azkenaldiko lan nabarmen batzuek etiketatze morfosintaktikora mugatzea hautagaiak identifikatzeko aurreprozesamendua (Inkpen eta Hirst, 2002; Dias, 2003). Seretanen ustez, ingelesaren morfologiaren eta sintaxiaren "sinpletasun erlatiboak" erraztu egiten du teknika oinarrizkoagoak erabiltzea, eta beste hizkuntza batzuetan (alemana, frantsesa...), teknika aurreratuagoak behar dira emaitza onargarriak lortuko badira.

Euskararen kasuan, badugu irudipena informazio sintaktikoak lagun liezaiokeela UF hautagaiak doitasun eta estaldura handiagoz erauzteari; batik bat, euskara aski ordena librekoa denez, kolokazioen osagaiak ondoz ondokoak izate hutsaz baino eraginkorra izan liteke mendekotasun-informazioaz baliatzea. Nolanahi ere, aztertu egin behar da zein garatze-mailataraino heldu den analisi sintaktiko azalekoa zein sakona egiteko teknologia, eta gure ikerkuntzan aplikatzeko moduan dagoen. Horretaz jardungo dugu lan esperimentalaren diseinua aurkeztean (V.2 atala).

III.4 Karakterizazio-atazak eta ebaluazioa

UF izateko hautagaiak erauzi ondoren, horiek idiomatikotasunaren arabera karakterizatzea izaten da hurrengo urratsa. Izan ere, erauzketan atzematen diren konbinazio hautagaien artean badira UF ez direnak, konbinazio libreak baizik. Bestetik, UF direnen artean ere denek ez dute idiomatikotasun-maila bera. Idiomatikotasunaren continuumaz eta UFen sailkapenaz mintzatu gara dagoeneko, eta ideia horiek eraman dute ikertzaileen komunitatea karakterizazioa bi ataza-motatara bideratzera, III.2 atalean UFen eskuratze-prozesuaren urratsak bereizi dituzunean aurreratu dugun bezala:

- **Ranking bidezko karakterizazioa**. Continuumaren ideian oinarrituta, karakterizazioaren helburua da hautagaiak idiomatikotasun-mailaren arabera ordenatzea, rankingak antolatzea.
- **Sailkapen automatikoaren bidezko karakterizazioa**. UFen sailkapenean bereizitako kategorietan oinarrituta, karakterizazioaren hel-

burua da erauzitako hautagaiak automatikoki sailkatzea; esaterako, esapide idiomatikoak, kolokazioak eta konbinazio libreak bereiziz.

Horiek dira, hain zuen ere, literaturan deskribatu diren bi karakterizazio-ataza nagusiak (Evert, 2005: 25-26; Pecina, 2009: 26), jarraian jorratuko ditugunak.

III.4.1 Ranking bidezko karakterizazioa

UF hautagai multzo bat izaki, hurrengo III.5 atalean deskribatuko ditugun tekniketako baten bidez hautagai bakoitzaren idiomatikotasunaren propietate jakin baten kuantifikazioa lor dezakegu, eta neurketa horren emaitzen arabera ordenatu hautagaiak. Rankingean behera joan ahala, espero duguna da idiomatikotasun-maila ere apalduz joatea (Heid, 2008: 351). UFak zenbat eta gorago egon rankingean, hainbat eta hobea da erabilitako teknika.

Ranking horiek erabiltzeko bi modu nagusi daude (Evert, 2008: 6):

- Erauzle automatikoaren emaitzak, edo horien parte bat, aditu batzuek aztertzea, hautagaia benetako UFa den ala ez ebazteko. Erauzketa erdiautomatikoa da, beraz, eta oso prozedura erabilia da, ikerkuntzan ez ezik (erauzlearen ebaluazioa egiteko), kolokazioak erauzteko tresna eta software komertzialetan ere. Lexikografian eta terminologian izan dute aplikazioa batez ere tresna horiek. Emaitzak osorik aztertzen ez direnean, irizpideren bat behar da aztergaitzat hartuko direnak aukeratzeko (atari bat, kopuru bat...).

- Aplikazio batzuetan, interesgarria izan daiteke ranking jarraitu horretako hautagaien bereizketa, eskuarki bitarra, sortzea, rankingaren posizio edo neurriaren atari-balio batetik gorakoak UFtzat joz, eta hortik beherakoak baztertuz. Sailkpenaren antzeko ataza da, beraz. Aplikazio automatikoetan da ideia hori aplikagarria, emaitzak orraztu gabe erabiltzen baitira.

Batean zein bestean, atari egokienaren auzia dugu. Horren inguruko ikerketa handia egin arren, gai irekia da oraindik, eta ez dirudi orotarako atari bat ezartzeko funtsik dagoenik (Krenn, 2000: 125; Inkpen eta Hirst, 2002). Gainera, aplikazioaren arabera, atari zorrotza edo lasaia komeni daiteke. Izan ere, atari zorrotzak lasaiak baino hautagai gutxiago hautatzen dituenez, doitasuna lehenesten du, estalduraren kaltetan, eta alderantziz. Aplikazio guztiz automatikoetan, doitasuna hobestea komeni da, oro har, baina esperimentalki ezarri behar izaten da zein den aplikazioaren emaitza onenak

dakartzan atari-balioa. Proiektu lexikografiko edo terminologikoetan, berriz, hobe da, proiektuan inberti daitezkeen baliabideen arabera, estaldura handia izatea.

III.4.2 Sailkapen automatikoaren bidezko karakterizazioa

UFen erauzketan, ranking bidezko emaitzak lortzea izan da prozedura erabiliena urtetan, eta lan handia egin da emaitza onenak dituen neurria aurkitu nahian. Baina idiomatikosuna gertakari konplexua da, zenbait propietate edo ezaugarriren konbinazioa, eta, beraz, zentzuzkoa da pentsatzea propietate horien neurketak konbinatuz ere ekin lekiokeela karakterizazioari, edota emaitzak hobetu ere egin litezkeela. Hori ez ezik, UF-kategoria desberdinak daude, eta ranking bidezko sistemak sailkapena egiteko dituen berezko mugak kontuan izanik, azken hamarkadan esperimentazio-lan handia egin da idiomatikotasunaren karakterizazioa ikasketa automatikoaren bidezko sailkatze-atazatzat planteatuz.

Horrelako sistemetan, UF hautagai bakoitzaren neurketak atributu dira (*feature*), eta sailkatze-algoritmoak eskuz sailkatutako adibide-multzo bat (*training set* edo ikaste-multzoa) erabiltzen du trebatzeko, hau da, ezagutza horietatik UFak sailkatzen ikasteko. Ondoren, beste adibide sailkatu multzo bat (*test set* edo test-multzoa) erabiltzen da sailkatzailea ebaluatzeko (ebaluazio horretan, jakina, sailkatzaileak ez du ikusten adibide bakoitza zein sailetakoa den).

UFen kasuan, lehen saioak, III.10 atalean ikusiko dugunez, idiosinkrasia estatistikoaren neurriekin egin ziren. Sistema horietan, eskuarki sailkapena bitarra izaten da, hau da, UF hautagaiak bi kategoriatan banatzen dira: UFtzat jo direnak eta konbinazio libretzat jo direnak. Geroago, idiomatikotasunaren beste zenbait propietateren neurketak ere konbinatu dira, eta horrelakoetan, UF-kategoriak bereizteko hainbat esperimentu egin dira (arruntena, *esapide idiomatikoak* / *kolokazioak* / *konbinazio* libreak bereiztea).

III.4.3 Karakterizazio automatikoaren ebaluazioa

Aurreko bi atazetan lortzen diren emaitzak izaera desberdinekoak dira, baina, karakterizazio-sistemaren ebaluazioa egiteko, bistan da hautagai bakoitza ebaluatuta eduki behar dugula; hautagaiaren benetako izaera zein den jakiteko modua antolatu behar da: UF den ala ez, eta, UF bada, zein kategoriatakoa.

Horretarako, lehenik erabaki behar dugu zer ebaluatuko dugun, emaitza guztiak edo horien parte edo lagin bat; hau da, ebaluatzeko hautagaien multzoa zein den. Bestetik, erreferentzia bat behar dugu, *gold standard*tzat erabiliko duguna. Azkenik, ebaluazioa egiteko neurri-sistema bat behar da, emaitza erreferentziarekin konparatutakoan, ebaluazioaren emaitza emango duena. Hurrengo hiru ataletan landuko ditugu ebaluazio-metodo bat eratzerakoan zehaztu beharreko alderdiak.

III.4.3.1 Ebaluazio-lagina

Karakterizazio-sistemaren ebaluazioa egiteko erabiliko ditugun hautagaien multzoa da. Aukera hauek erabil ditzakegu:

- Erauzketaren emaitza osoa. Aski multzo handiak izaten dira; horregatik, gutxitan erabiltzen dira ebaluazioan, nolaz eta erreferentziatzat ez dugun erabiltzen aurrez eratutako baliabide lexikal bat.
- Azpimultzo bat, irizpide hauetako baten arabera sortua:
  - Maiztasun-atari bat edo idiomatikotasun-neurriaren atari-balio bat gainditzen ez duten konbinazioak baztertuz osatutako lagina. Ataria teorikoki (esangura-testak eta p balioak erabiliz) edo esperimentalki ezar daiteke, baina, lehen atariaz adierazi ditugun arazoez gain, Evertek dio, neurriaren arabera, tamaina desberdineko laginak sortzen direla sistema honekin, eta zaila dela neurrien portaera kondizio horietan konparatzea (Evert, 2005: 138).
  - *n-best list* edo rankingeko lehen n hautagaien zerrenda. Aurreko puntuan azaldutako arazoaren irtenbidetzat proposatu da prozedura hau (Krenn, 2000: 125; Evert, 2005: 138-139). Hala ere, ebaluazio esanguratsu eta fidagarriak egiteko, eskulan handia eskatzen duten laginak etiketatu behar dira, eta hori egitea kostu handikoa izaten da.
  - Ausazko lagin bat: eskulanaren kostua gutxitu nahian, emaitza guztien ausazko lagin baten bidezko ebaluazio-prozedurak proposatu dira (Evert, 2005: 159-162). Everten sisteman, hautagai guztien % 10eko lagina etiketatzen da eskuz, eta horien rankingetik estimatzen erauzketa osoaren portaera (doitasuna eta estaldura). Arazoa, Evertek berak aitortzen duenez, $n \leq 1\,000$ tartea da, doitasunaren estimazioa ez baita fidagarria (estimazioa egiteko, 100 hautagai etiketatu inguru baino ez daude tarte horretan).

III.4.3.2 Erreferentzia edo *gold standard*a

Ebaluazioan erabiliko ditugun hautagaien multzoa zehaztu ondoren, hautagai bakoitza UF den ala ez jakiteko prozedura antolatu behar da. Hauek dira aukera ohikoenak:

- Kanpo-baliabide lexikal bat (hiztegiak, datu-base lexikalak, WordNet). Abantaila bistakoa da: eskulana aurreratu egiten da. Eragozpen handia da, ordea, horrelakoen estalduraren mendeko garela. Hain zuen ere, autore batzuek aitortu dute eta ez dela arraroa erauzketaren emaitzetako UF asko asko hiztegi-erreferentzian falta izatea (Korhonen et al., 2006: 1018). Horrekin lotutako beste desabantaila bat da horrelako erreferentzia batek ez digula aukerarik ematen ebaluatu nahi dugun neurriak gure corpus jakinean duen portaera ezagutzeko; esaten digun bakarra da erauzketako zein hautagai ageri den erreferentzian (Krenn et al., 2004).

- Ebaluazio-laginaz mintzatzean aipatu dugun erauzketaren emaitzaren azpimultzo hura bera, eskuz etiketatuta. Etiketa-motak:

  – Bitarra: UF bai/ez

  – UF-mota edo -kategoria

  – Idiomatikotasun-eskala bat

  Ranking-bidezko atazan, sistema bitarra da erabiliena, baina sistema horrek ez du balio UF-kategoriak bereizteko edo idiomatikotasun-continuuma birsortzeko. Dena den, badira eskala jarraitu bat erabiltzen dutenak erreferentziako hautagaiak etiketatzeko (McCarthy et al., 2003; Biemann eta Giesbrecht, 2011). Sailkatze-atazan, sistema bitarra erabil daiteke, baina, batez ere, UF-kategoriak bereizten dituena da interesgarriena, eta, horretarako, hiru kategoria behar dira gutxienez.

  Erreferentziaren sailkatze-lana aditu-multzo batek egin ohi du. Autore askok azaldu dute UF hautagaiak etiketatzea ez dela egiteko erraza, eta subjektibotasunaren arriskuak minimizatzearren, gomendatzen da (Krenn et al., 2004), batetik, etiketatzaileei UF kontzeptuaren eta UF-kategorien inguruko argibide zehatzak ematea, eta, bestetik, etiketatzaileen arteko adostasuna (ITA - *inter-tagger agreement*) kontrolatzea, arrazoizko gutxieneko adostasuna bermatzeko. ITA neurtzeko, arlo honetan ohikoena κ neurriak erabiltzea da, Cohen κ bi etiketatzaile direnean, eta Fleiss κ bi baino gehiago direnean (Gwet, 2012).

Eskuzko *gold standard*ak konpontzen du nolabait hiztegien estaldura-ren arazoa, baina desabantaila da eskulan handia egin behar dela. Horregatik, erreferentzia, eskulana gehiegizkoa izan ez dadin, ezin da oso handia izan, baina bai estimazio fidagarria egiteko behar den bezain handia (Pecina, 2009: 50). Bestetik, dagoeneko adierazi dugu eskulana gutxitzeko asmoz ausazko laginen ebaluaziotik ebaluazio-bilduma osoaren doitasuna eta estaldura estimatzeko teknikak proposatu direla.

III.4.3.3 Metrika

Ranking-atazaren kasuan, sistema erabiliena da UFen propietate baten neurketaren bidez eratutako rankinga erreferentzia bitar batekin konparatzea. Rankingeko lehen n itemak (*n-best list*) "hautatuak" edo positiboak dira (P). Horietako bakoitza erreferentzian dagoen begiratuz, jakin dezakegu egiazko positiboa den (*true positive*) ala positibo faltsua den (*false positive*). Beraz, rankingeko lehen hautatuen multzo bat edukita, badakigu zein den egiazko positiboen kopurua (TP) eta positibo faltsuen kopurua (FP). Era berean, jakin dezakegu *n-best list* bakoitzean erreferentziako zenbat item ez diren agertu; horiek negatibo faltsuak dira (*false negatives*, FN). Azkenik, laugarren multzoa da *n-best list*ean agertzen ez diren eta UF ez diren itemak; egiazko negatiboak dira (*true negatives*, TN).

Horien kopuruak kontingentzia-taula edo konfusio-matrize batean antolatu ohi dira (III.1 taula).

	erreferentzian	¬erreferentzian
hautatuak (P)	egiazko positiboak (TP)	positibo faltsuak (FP)
¬hautatuak (N)	negatibo faltsuak (FN)	egiazko negatiboak (TN)

III.1 Taula: IR sistema baten irteeraren kontingentzia-taula edo konfusio-matrizea.

Informazio hori erabiliz neur dezakegu erauzte-sistemaren eraginkortasuna. Kasu horretan, metrika erabilienak dira Informazioaren Berreskurapenaren (IR) arloan estandarrak diren doitasuna (P), estaldura (R) eta horien konbinazioa den F neurria (Baeza-Yates eta Ribeiro-Neto, 1999). Jarraian emango ditugu neurrion adierazpenak. Lehenik, doitasuna:

$$P = \frac{TP}{TP + FP} \tag{III.1}$$

Beraz, doitasuna da sistemak egiazkotzat itzultzen dituenetatik, benetan egiazkoak direnen proportzioa.

$$R = \frac{TP}{TP + FN} \tag{III.2}$$

Estaldura da sistemak itzuli behar lituzkeen egiazko guztietatik itzuli dituen egiazkoen proportzioa.

Bi neurri horien arteko erlazio zuzenik ez dagoen arren, azterketa enpirikoetan ikusi da alderantziz erlazionatuta daudela: sistemak detektatutako elementuen kopurua handitzean (estaldura handitzen bada), doitasuna gutxitzen da, eta alderantziz (Buckland eta Gey, 1994). Biak neurri batean konbinatzeko, F neurria proposatu da, Van Rijsbergenek (1979) sortutako E neurriaren aldaera dena. Doitasun eta estalduraren batezbesteko haztatua da:

$$F_\beta = (\beta^2 + 1) \cdot \frac{P \times R}{(\beta^2 \times P) + R} \tag{III.3}$$

Normalean, $\beta = 1$ erabiltzen da, doitasunari eta estaldurari pisu bera emanez:

$$F_1 = 2 \cdot \frac{P \times R}{P + R} \tag{III.4}$$

Baina erreferentzia-lagina bi kategoriatan baino gehiagotan sailkatua dugunean, edo idiomatikotasun-eskala jarraitu bat dugunean, heinen korrelazio-koefizienteak erabili ohi dira (*rank correlation coefficient*), ranking ideala eta esperimentala konparatzeko. UFen erauzketaren arloan, erabilienak Spearman ρ eta Kendall τ dira (Fredricks eta Nelsen, 2007). Horrelakoetan, garrantzitsua da rankingetako ordena-berdinketak (*ties*) egoki kudeatzea; horretarako, neurriek berdinketak daudenean erabili beharreko bertsioak dituzte (Gibbons, 1993: 7).

Sailkatze-atazan, ikasketa automatikoan estandarrak diren neurriak erabili ohi dira. Batetik, ranking-atazan bezala, P, R eta F neurria erabiltzen dira, lehen biak ataza honetarako honela egokiturik:

$$P = \frac{\text{kategoria batean zuzen sailkatuak}}{\text{kategoria horretan sailkatuak}} \tag{III.5}$$

$$R = \frac{\text{kategoria batean zuzen sailkatuak}}{\text{kategoria horretako guztiak}} \tag{III.6}$$

Sailkapenean bi kategoria baino gehiago dagoenean, kategoria bakoitzerako 2×2 kontingentzia-taulak egiten dira, eta, ebaluazio-neurri globalak kalkulatzeko, bi estrategia erabil daitezke:

- *Microaveraging* (mikrobatezbestekoa): kontingentzia-tauletako mota bereko balioen batura erabiltzen da ebaluazio-neurrien kalkuluan. Ez da beharrezkoa kategoria bakoitzeko emaitzak kalkulatzea.

- *Macroaveraging* (makrobatezbestekoa): kategoria bakoitzaren ebaluazio-neurriak kalkulatu, eta gero horien batezbestekoa egiten da.

Lehenak pisu bera ematen die kategoria guztiei, eta bigarrenak, pisu bera instantzia guztiei (Manning eta Schütze, 1999: 577). Mikrobatezbestekoa da erabiliena, baina arazo bat du: kategoria handietan lortzen diren emaitzen mendekoa da; hau da, instantzia-kopurua kategoriaka desorekatua bada, kategoria minoritarioetan egindako errore sistematikoak ez dira penalizatzen, eta gerta daiteke kategoria horiek gure interesekoak izatea. Muturrera eramanda, kategoria bateko instantziak % 90 edo gehiago badira, oinarri-lerroa (denei kategoria hori esleitzea) gainditzea oso zaila izan daiteke. Alde horretatik, makrobatezbestekoa kategoria guztietan lortzen den kalitatearen neurri adierazgarriagoa da.

Ohiko beste neurri bat zehaztasuna da (*accuracy*). Zuzen sailkatutako instantzien ehunekoa da:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{III.7}$$

III.5 Idiomatikotasunaren propietateak neurtzeko estrategiak

Idiomatikotasunaren kontzeptu zabala definitu dugunean, lau osagai edo propietate bereizi ditugu. UFen karakterizazioa egiteko garatu diren teknikak UFen ezaugarri baten edo batzuen detekzioan edo neurketan oinarritzen dira.

Beraz, idiomatikotasunaren adierazgarri diren aipatu propietate edo ezaugarriak fenomeno edo behagai jakinen bidez karakterizatu edo neurtu ohi dira. Hauek dira ikerkuntzan gehien erabili diren prozedurak:

- Idiosinkrasia estatistikoa (instituzionalizazioa): konbinazio hautagaiaren osagaien **agerkidetza**-informazioa (*coocurrence*) prozesatzen da, elkartze-neurri estatistikoak erabiliz.
- Konposizionaltasun-maila (semantikoa): konbinazioaren eta haren osagaien **antzekotasun distribuzionala** (*distributional similarity*) edo testuinguruen arteko antza neurtzen da.
- Finkotasuna
  - Morfosintaktikoa: konbinazioaren portaera morfosintaktikoak egitura bereko konbinazioen **batez besteko portaera morfosintaktikoarekiko duen distantzia**ren bidez neurtzen da.
  - Lexikala: konbinazioaren osagaien **ordezkagarritasuna** (*substitutability*) neurtzen da, osagaien ordezkotzat haien sinonimoak edo kuasisinonimoak erabiliz, eta sortzen diren konbinazio-aldaeren propietate estatistikoak jatorrizko konbinazioaren propietateekin konparatuz.

Hurrengo lau ataletan, idiomatikotasunaren propietate horietako bakoitzaren neurketa-teknikak azalduko ditugu.

III.6 Idiosinkrasia estatistikoaren neurketa

II.1.2.1 atalean azaldu dugu idiosinkrasia estatistikoa eta instituzionalizazioa erlazionatuta daudela, erabat kontzeptu baliokideak ez badira ere. Idiosinkrasia estatistikoaren funtsezko ideia da konbinazio baten maiztasuna nolabait nabarmena dela; bestetik, esan dugu nabarmentasun horren ideia zehatzagoa dela konbinazio baten maiztasuna handiagoa izatea konbinazio horri ausaz legokiokeen maiztasuna baino, hau da, hitzak konbinatzeko arauak eta osagai bakoitzaren maiztasuna kontuan harturik sortuz gero izango lukeen maiztasuna baino (Church eta Hanks, 1990: 23). Bestela esanda, osagaien artean badela korrelazio, asoziazio edo elkarrekin agertzeko joera estatistikoki esanguratsua.

Elkartze-maila hori zehazteko erabiltzen diren neurriei *lexical association*

measures (AM) edo *hitzen elkartze-neurriak* izendapena ematen zaie maiz (Evert, 2005: 20; Urizar, 2012: 171).

III.6.1 Agerkidetza-datuak eta kontingentzia-taulak

Elkartze-neurriek hitzen agerkidetzaren informazioa erabiltzen dute, hau da, konbinazioaren osagaiak elkarrekin agertzen diren maiztasuna eta bata bestea gabe agertzen diren maiztasunak. Informazio hori *kontingentzia-taula* batean errepresentatzen da. Esaterako, corpus batetik N bigrama-kopurua erauzi badugu, hau da (A, B) bigrama jakin bakoitzerako kontingentzia-taula:

	$w_2 = B$	$w_2 \neq B$	
$w_1 = A$	O_{11}	O_{12}	$= R_1$
$w_1 \neq A$	O_{21}	O_{22}	$= R_2$
	$=C_1$	$= C_2$	

N bigrama-kopurua bada:

$$O_{11} + O_{12} + O_{21} + O_{22} = N \tag{III.8}$$

Bestetik:

- $O_{11} + O_{12} = R_1$: $(A, \text{-})$ bigrama-kopurua, edo A hitza duten bigramen kopurua
- $O_{11} + O_{21} = C_1$: $(B, \text{-})$ bigrama-kopurua, edo B hitza duten bigramen kopurua
- $O_{21} + O_{22} = R_2$: $(\neq A, \text{-})$ bigrama-kopurua, edo A hitza ez duten bigramen kopurua
- $O_{12} + O_{22} = C_2$: $(\neq B, \text{-})$ bigrama-kopurua, edo B hitza ez duten bigramen kopurua

R_1 eta C_1 aldagaiei *maiztasun marjinal* deitzen zaie (*marginal frequencies* edo *joint frequencies*). Kontuan izan behar dugu maiztasun marjinalak ez direla A edo B-ren corpuseko agerpen-kopuru totalak, baizik eta, hurrenez hurren, lehen osagaitzat A duten bigramen agerpen-kopurua eta bigarren

osagaitzat B duten bigramenena. Laukote honi bigrama baten *maiztasun-sinadura* esaten zaio (*frequency signature*):

$$f(w_1, w_2), f_1(w_1), f_2(w_2), N =: (f, f_1, f_2, N)_{(w_1, w_2)}, \tag{III.9}$$

non:

$$\begin{aligned} f &= O_{11} \\ f_1 &= O_{11} + O_{12} = R_1 \\ f_2 &= O_{11} + O_{21} = C_1 \\ N &= \sum_{ij} O_{ij} \end{aligned}$$

Bigramak sortzen direnean, erosoagoa da maiztasun-sinaduran dauden balioak zenbatzea, eta horietatik ateratzea kontingentzia-taulako maiztasunak.

Esaterako, hau da *egunkarian irakurri* bigramaren maiztasun-sinadura tesi-lan honetan erabili dugun corpusean (ikus V.1 atala)[2].

$$(f, f_1, f_2, N)_{(egunkarian, irakurri)} = (43,\ 406,\ 4\,068,\ 6\,136\,897)$$

Beraz, dagokion kontingentzia-taula:

	$w_2 = irakurri$		$w_2 \neq irakurri$	
$w_1 = egunkarian$	43	+	363	= 406
	+		+	
$w_1 \neq egunkarian$	4 025	+	6 132 466	= 6 136 491
	= 4 068		= 6 132 829	N = 6 136 897

Bestela adierazita:

[2] Zehazki, VI.3 atalean deskribatuko dugun $w = \pm 1$, $f \geqslant 30$ erauzketaren emaitzen araberako datuak dira.

N	bigrama-kopurua	6 136 897
O_{11}	*egunkarian irakurri* bigramaren agerpen-kopurua	43
R_1	*egunkarian* hitza duten bigramen kopurua	406
C_1	*irakurri* hitzaren duten bigramen kopurua	4 068
R_2	*egunkarian* hitza ez duten bigramen kopurua	6 136 491
C_2	*irakurri* hitza ez duten bigramen kopurua	6 132 829

III.6.2 Agerkidetzaren eredu estatistikoa eta ausazkotasuna

Esan dugunez, AMetan hitz-konbinazio baten maiztasuna eta osagaiek UF osatuko ez balute espero litekeen, hau da, itxarondako agerkidetza-maiztasuna konparatu ohi dira (Heid, 2008: 350).

Lehena behaketaren emaitza da, eta planteatzen duen galdera da behaketa egiteko erabili dugun lagina zenbateraino den populazio osoaren adierazgarria. Galdera horren erantzuna agerkidetzaren informazioa eman digun corpusaren izaeran eta eratzeko metodologian dago.

Bigarrena, berriz, estimazio baten emaitza da, laginaren banaketaz dugun suposizioan oinarritua. Hitz-konbinazioa UFa ez bada, ez du idiosinkrasia estatistikorik, eta, ausazko laginketa-eredua asumituz gero, bi hitzak elkarren ondoan gertatzea ausazko gertakaria da, hitz bakoitza bere probabilitatearen arabera agertzen da testuan, bata bestetik independenteki. *Independentziaren hipotesi nulua* deritzo (H_0) bi behagai edo fenomeno elkarren arteko loturarik gabeak direla suposatzeari. (w_1, w_2) bigramaren osagai biak batera gertatzeko itxarondako probabilitatea bien banako probabilitateen biderkadura da:

$$P(w_1, w_2) = P(w_1) \cdot P(w_2) \tag{III.10}$$

Hori kalkulatzeko, egiantz handieneko estimatzaileak (*maximum-likelihood estimator*) erabil daitezke osagaien probabilitateetarako (Evert, 2005: 49-52), hau da, maiztasun erlatiboak. Beraz, aurreko taulako (A, B) bikotearen kasuan:

$$P(A, B) = P(A) \cdot P(B) = \frac{R_1}{N} \cdot \frac{C_1}{N} = \frac{R_1 C_1}{N^2} \tag{III.11}$$

Hortaz (A, B) bikotearentzat ausaz espero daitekeen maiztasuna edo agerpen-kopurua:

$$E_{11} = \frac{R_1 \cdot C_1}{N}$$

Hurrengo taulan, kontingentzia-taulako gainerako ausazko maiztasunak daude:

	$w_2 = B$	$w_2 \neq B$
$w_1 = A$	$E_{11} = R_1C_1/N$	$E_{12} = R_1C_2/N$
$w_1 \neq A$	$E_{21} = R_2C_1/N$	$E_{22} = R_2C_2/N$

Esate baterako, *egunkarian irakurri* konbinaziorako:

	$w_2 =$ *irakurri*		$w_2 \neq$ *irakurri*	
$w_1 =$ *egunkarian*	0,27	+	405,73	= 406
	+		+	
$w_1 \neq$ *egunkarian*	4 067,73	+	6 132 423,27	= 6 136 491
	= 4 068		= 6 132 829	N = 6 136 897

AMek behatutako maiztasun-datuak (O) eta independentzia-hipotesi nuluaren arabera espero litezkeen maiztasun-datuak (E) konparatzen dituzte, eredu estatistiko desberdinetan oinarrituta, independentzia-hipotesi horretatik modu esanguratsuan aldentzen diren konbinazioak detektatzeko.

Hala ere, zenbait ikertzailek ohartarazi dutenez (Evert, 2005: 57-59; Bouma, 2010; Ramisch, 2012: 45) ausazkotasunean oinarritutako independentzia-irizpidea ez da hizkuntzaren kasuan aplikagarria. Everten hitzetan (Evert, 2008: 1254):

> ≪The null hypothesis of independence is extremely unrealistic. Words are never combined at random in natural language, being subject to a variety of syntactic, semantic and lexical restrictions.≫

Adibide aipatuenetakoa ingelesezko *the the* bigrama da (Baayen, 2001). Ingelesezko gramatikak ia ezinezkoa egiten du horrelako segida, baina independentziaren suposizioan, konbinazio horren probabilitatea ez da nulua, aski handia baizik (Baayenen arabera, 0,0036; horrek esan nahi du 300 hitzez behin espero dezakegula segida hori gertatzea). Beraz, independentziaren hipotesiak itxarondako probabilitateak gainestimatzea dakar berekin,

eta zaildu egiten du oso maiztasun handiz agertzen ez diren konbinazioen elkartze-neurria nabarmentzea.

Evertek iradokitzen du irtenbide posible bat dela hitz-konbinazioen murrizketak kontuan hartuko lituzkeen hipotesi nulu errealistago bat zehaztea, baina aitortzen du ikerketa hori hastapenetan dagoela.

Boumak (2010) argudiatzen du berez kolokazionalak ez diren elkartze edo asoziazio batzuk nabarmendu egin daitezkeela benetako kolokazioak direnen gainetik. Ildo horretatik, hau da, independentziaren asuntzioaz haraindi joan nahian, konbinazioaren itxarondako probabilitate "informatuagoa" estimatzeko prozedura landu nahi du, eta horretarako, *Aggregate Markov Models* teknika erabiltzea proposatu du. Boumak berak aitortzen duenez, proposatutako metodo berriaren eraginkortasuna aldakorra da, eta lehen esploraziotzat aurkezten du bere lana.

Itxarondako probabilitatea definitu dugun moduan estimatzearekin erlazionatutako beste arazo bat oso maiztasun txikiko konbinazioena da. Horrelakoak agerkidetzaren analisiaren bidez atzematea erronka bat da, eta autore askok gomendatzen dute maiztasun minimo bat ezartzea konbinazio bat hautagaitzat jotzeko. Izan ere, maiztasun batetik behera, estatistikoki esanguratsuak diren konbinazioak ezin dira zori hutsez eratutako konbinazioetatik modu fidagarrian bereizi. Arazoa bide da ez dagoela araurik edo algoritmorik atari horren balioa ondorioztatzeko (Ramisch, 2012: 44-45). Nolanahi ere, Evertek erakutsi du (Evert, 2005: 132-133) inferentzia estatistikoa ezinezkoa dela *hapax* eta *dis legomena* gertakarientzat ($f = 1, 2$), batez ere kuantizazio-efektuen ondorioz, eta, maiztasun-atari bat ezartzeko arrazoi teorikoak daudela dio. $f < 3$ konbinazioak sistematikoki kontuan ez hartzea aldarrikatzen du, eta $f \geqslant 5$ konbinazioak soilik kontsideratzen ditu kuantizazio-efektuen eraginetik libre.

III.6.3 Elkartze-neurriak (AM)

UFen eta, bereziki, kolokazioen erauzketan erabili diren AMen multzoa oso da ugaria. Hainbat neurri esperimentatu eta ebaluatu dira, egiteko honetarako neurri eraginkorrena aurkitu nahian. Atal honetan, UFen erauzketan gehien erabili diren AMak aurkeztuko ditugu.

AMen tipologia egitean, Evertek, estatistikako oinarri matematikoa kontuan harturik, bi multzo nagusi hauek bereizten ditu (Evert, 2008: 22):

- Asoziazioaren **esangura** (*significance*) neurtzen dutenak. Neurri hauek galdera honi erantzun nahi diote: "zenbaterainoko ebidentzia dugu hitzen artean badela asoziazio positibo bat?" Horri erantzuteko, hau

neurtzen dute: zenbateraino den ez-probablea hitzak independente direla dioen hipotesi nulua. *Hipotesi-egiaztatze*aren arloko neurriak dira, beraz. Multzo zabala da hau, Evertek hiru azpimultzotan banatzen duena[3]:

  - egiantz-neurriak (esaterako, Poisson-Stirling). Ez dira oso erabiliak izan kolokazio-erauzketan.
  - hipotesi-test zehatzak (esaterako, Fisherren test zehatza).
  - hipotesi-test asintotikoak (esaterako, z neurria, t neurria, Pearsonen χ^2 testa eta egiantz-arrazoiaren logaritmoa).

- Asoziazioaren **efektuaren tamaina** (*effect size*) neurtzen dutenak. Hauen galdera da "zenbaterainokoa da hitzen arteko erakarpena edo elkartze-maila?" (*association strength*); horri erantzuteko, behatutako maiztasuna itxarondako maiztasunetik zenbateraino urruntzen den neurtzen dute. Multzo honetan sartzen dira, besteak beste, μ balioa, momioen arrazoia (*odds-ratioa*) eta Dice koefizientea.

Nahiz eta sailkapen xeheago batean aparteko sailean antolatu (Evert, 2005: 76-77), Evertek (2008) Informazioaren Teoriatik datozen neurriak aurreko eskeman sartzen ditu. Elkarrekiko informazioaren kasuan (*mutual information*, MI), elkarrekiko informazio puntuala izeneko aldaera (*pointwise mutual information*, PMI) efektuaren tamainaren saileko μ balioaren baliokidetzat jotzen da.

Multzo horietatik kanpo, zenbait neurri heuristiko geratzen dira, oinarri matematiko zehatzik ez izan arren, UFen erauzketarako proposatu eta erabili direnak. Nagusiak dira MI^k familiakoa (ezagunena MI^3) eta maiztasuna bera (f).

Jarraian, UFen erauzketan gehien erabili diren AMen deskribapen zehatzagoa egingo dugu: z neurria, t neurria, Pearsonen χ^2 testa, egiantz-arrazoiaren logaritmoa (LLR), Fisherren test zehatza, elkarrekiko informazio puntuala (PMI), MI^3 eta f.

z neurria (*z-score*)

Kolokazioak erauzteko erabilitako lehen neurrietako bat da z neurria (Berry-Rogghe, 1973). Banaketa normalean oinarritua dagoen alde bateko test parametrikoa da. Behatutako balioaren (x) eta populazioaren batezbestekoaren (μ) arteko diferentzia neurtzen du, σ desbideratze estandarrarekiko:

[3] http://www.collocations.de/AM/index.html

$$z = \frac{x - \mu}{\sigma} \tag{III.12}$$

z neurrian, laginaren banaketa binomial diskretuaren hurbilpena egiten da, banaketa normal jarraitu bat erabiliz. Kondizio horietan, honela kalkula daiteke, behatutako eta itxarondako maiztasunen balioetatik (Evert, 2005: 80):

$$z = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}} \tag{III.13}$$

Banaketa normalaren bidezko aipatutako hurbilpena da z neurriaren desabantailetako bat, hurbilpena itxarondako maiztasun handietarako bakarrik baita fidagarria. Itxarondako maiztasuna txikia denean ($E < 1$), oso ez-zehatza da. Hurbilpena hobetu egin daiteke, Yatesen jarraitutasun-zuzenketak aplikatuz[4].

Hala ere, z neurria populazioaren parametroak ezagunak direnean soilik dago definituta, eta Manning eta Schützeren (1999) arabera, bariantza ezaguna denean soilik erabili behar litzateke.

t neurria (*t-score*)

z neurria bezala, alde bateko test parametrikoa da, lagina banaketa estandarra duen populazio batetik ateratakoa dela asumitzen duena. Laginaren batezbestekoaren ($\overline{x}$) eta hipotesi nuluaren arabera estimatutako populazioaren batezbestekoaren (μ) arteko diferentzia neurtzen du, errore estandarrarekiko (Pedersen, 1996):

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}}, \tag{III.14}$$

non σ^2 laginaren bariantza den, eta N populazioaren tamaina.

$\overline{x}$, μ eta σ^2 estimatzeko, prozedura hau proposatu da (Manning eta Schütze, 1999: 154): hipotesi nuluan, asumi daiteke laginean bigrama jakin bat behatzea Bernoulli saio baten emaitza bezalakoa dela. Horretara, $\overline{x}$-ren balioa bigramaren probabilitatea da: $P(w_1, w_2)$ (egiantz handieneko estimatzailea erabiliz, behatutako maiztasun erlatiboa, O_{11}/N); σ^2-ren balioa: $p(1 - p)$; eta μ-rena, azkenik, bigramaren itxarondako probabilitatea: $P(w_1) \cdot P(w_2)$ edo $E_{11} = R_1 C_1 / N$. Bigrama askotarako p txikia dela argudiatuta, aipatu autoreek $\sigma^2 = p(1 - p) \approx p$ hurbilketa egiten dute.

[4] http://www.collocations.de/AM/index.html

Orduan, froga daiteke (Seretan, 2011: 38) t neurriaren balioa honela kalkula daitekeela behatutako eta itxarondako maiztasunen balioetatik:

$$t = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}} \tag{III.15}$$

t neurria Church et al.-en (1991) lanean proposatu zen estreinakoz kolokazio-erauzketan aplikatzeko. Student t testaren bertsio bat erabili zuten, bariantzaren estimazioa Manning eta Schützek (1999) adierazitako modu berean egiten duena.

Zenbait autorek zalantzan jarri dute t neurria agerkidetza neurtzeko proposatutako eran aplikatzearen egokitasuna, bateraezinak bailirateke oinarrian duen banaketa normala eta laginaren bariantzaren estimazioa egiteko proposatu den banaketa binomiala (Evert, 2005: 82-83). Hala eta guztiz ere, Evertek berak ohartarazten gaitu t neurriak emaitza konparatibo onak izan dituela hainbat ikerlanetan (Evert, 2008: 36).

Pearsonen χ^2 edo khi karratuaren testa (*chi-squared test*)

Bi aldeko test ez-parametrikoa da (ez du banaketa jakin bat asumitzen). Datu-multzo bat χ^2 banaketara zenbateraino hurbiltzen den egiaztatzeko erabiltzen da. Taula bateko datuetan behatutako balioak eta aldagaiak independente balira espero litezkeen balioak konparatzen ditu, adierazpen honen bidez:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{III.16}$$

X^2 neurria χ^2 banaketara asintokikoki hurbiltzen dela froga daiteke, hau da, datu-multzoa handia bada, X^2 neurriak χ^2 banaketa agertzen du.

Bigramen kasuan ($j = 2$), adierazpen hori hurrengo honen baliokidea dela froga daiteke (Seretan, 2011: 133):

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11}O_{21})(O_{12} + O_{22})(O_{21} + O_{22})} \tag{III.17}$$

III.13 eta III.15 ekuazioen erara adierazirik:

$$\chi^2 = \frac{N(O_{11} - E_{11})^2}{E_{11}E_{22}} \tag{III.18}$$

χ^2 testa kontingentzia-tauletako zutabe eta errenkaden arteko independentzia egiaztatzeko test estandarra da, eta, datuen banaketa normalaren

suposizioan oinarritua ez denez, teorian abantaila izango luke banaketa normalean oinarritutako testen aldean, t neurriaren aldean, esaterako. Oro har, χ^2 testaren emaitzak probabilitate handietarako egokiagoak direla onartzen da (Manning eta Schütze, 1999: 159; Korkontzelos, 2010), eta, teorikoki bederen, ataza-sorta zabalagorako egokia izan daiteke. Hala ere, maiztasun txikiekin (corpus txikiak, maiztasun txikiko hitzen konbinazioak), χ^2 testaren emaitzak okerragoak izatea aurreikusten da, eta z neurriaren antzeko alborapen-arazoak ditu. Horiek arindu nahian, Yatesen jarraitutasun-zuzenketa ezartzea proposatu da (Evert, 2005: 83).

Egiantz-arrazoia (*likelihood ratio*, LR)

Dunningek (1993) argudiatu du testuen analisian laginen banaketa normala dela edo horren hurbilketa dela suposatuz egiten diren kalkulu estatistikoek huts egiten dutela lagin txikiekin (Dunning, 1993: 2). Hain zuzen ere, gertakari arraroak konparatzean, testuetan horrelako gertakariak ugariak izatea litzateke test horiek huts egiteko arrazoia. Dunningek banaketa binomialean oinarritutako testa proposatu du: *likelihood ratio* edo egiantz-arrazoiaren testa.

Egiantz-arrazoia (λ) bi hipotesiren egiantzen zatidura da, eta haren logaritmoa (*log-likelihood ratio*, edo LLR):

$$\log \lambda = \log \frac{L(H_1)}{L(H_2)}, \tag{III.19}$$

non H_1 mendekotasunik ezaren hipotesia den, eta H_2, mendekotasunaren hipotesia.

Dunningek dioenez, -2log λ balioa X^2 banaketara hurbiltzen da asintotikoki, χ^2 testa baino doiago hurbildu ere. Dunningen adierazpenaren berridazketa baliokide batzuk proposatu dira; horien artean sinpleena hau da (Evert eta Krenn, 2001):

$$-2 \log \lambda = 2 \sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}} \tag{III.20}$$

Adierazpen hori beste honen baliokidea dela froga daiteke (Kilgarriff, 1996; Matsumoto eta Utsuro, 2000):

$$\begin{aligned}-2\log\lambda = 2[&O_{11}\log O_{11} + O_{12}\log O_{12} + O_{21}\log O_{21} + O_{22}\log O_{22}\\ &- (O_{11}+O_{12})\log(O_{11}+O_{12}) - (O_{11}+O_{21})\log(O_{11}+O_{21})\\ &- (O_{12}+O_{22}) - \log(O_{12}+O_{22}) - (O_{21}+O_{22})\log(O_{21}+O_{22})\\ &+ (O_{11}+O_{12}+O_{21}+O_{22}) - \log(O_{11}+O_{12}+O_{21}+O_{22})] \quad \text{(III.21)}\end{aligned}$$

Dunningez gain, zenbait autorek fidagarritasun handia aitortzen diote LLRri (Daille, 1994; Kilgarriff, 1996; Evert eta Krenn, 2001). Nolanahi ere, badira hori kuestionatzen duten lan esperimentalak. Esaterako, Krenn eta Evertek (2001) erakutsi dute, alemanezko PP-V egiturako konbinazioen erauzketan, egiantz-arrazoiarekin t neurriarekin baino emaitza txarragoak lortzen direla. Bestetik, Pecinaren (2010) ebaluazio-saio handian, LLRren emaitzak oso apalak dira (PMIrenak baino okerragoak, eta t neurriarenak baino hobeak).

Fisherren test zehatza (*Fisher's exact test*)

Pedersenek (1996) Fisherren test zehatzaren p balioa proposatu zuen AMtzat, aurreko hipotesi-test asintokikoak baino zehatzagoa dela argudiatuz. Test zehatzek bezala, p balioaren konputazio zehatza egiten du (lagin handietarako soilik balio duen test asintotikoen hurbilpena egin beharrean), eta horretarako banaketa hipergeometrikoan oinarritzen da. Egiazko test zehaztzat jo daiteke, eta matematikariak aski ados daude 2×2 kontingentzia-taulen esangura-test egokiena dela (Yates, 1984: 428). Hala ere, konputazionalki kostu handikoa da. Bestetik, (Manning eta Schütze, 1999: 159) egileek ohartarazten duenez, ez dago argi Pedersenen (1996) beraren emaitzetan Fisherren testak χ^2 testak baino portaera hobea duen. Azkenik, Evertek erakutsi du (Evert, 2005: 112) Dunningen egiantza-arrazoiaren logaritmoak Fisherren testaren p balioen oso hurbilpen doia lortzen duela, eta, horretan oinarrituta, uste du ez dagoela arrazoirik konputazionalki hain eskakizun handiak dituen Fisherren testa erabiltzeko (Evert, 2008: 31).

Elkarrekiko informazio puntuala (*pointwise mutual information*, PMI)

Berez, Informazioaren Teoriako kontzeptua da, eta Church eta Hanksek (1990) erabili zuten lehen aldiz kolokazio-erauzketan. PMIk (edo, UFen erauzketaren arloan, eskuarki MI) aldagai batek beste bati buruz ematen duen informazioa neurtzen du. Gure kasuan, corpusean w_1 hitza agertzeak bere inguruan w_2 beste hitz bat agertzeari buruz ematen duen informazioa

da. Beraz, esan daiteke PMIk hitzen arteko independentzia neurtzen duela. Adibidez, badakigu i posizioan *parte* hitza agertzea eta i+1 posizioan *hartu* hitza agertzea ez dela ausazko gertakaria; hartara, *parte* hitza agertzeak adierazten digu hurrengo posizioan *hartu* hitza agertzeko halako probabilitatea dagoela (baldintzazko probabilitatea), *hartu* hitza testuko edozein posiziotan agertzeko dagoen probabilitatea baino handiagoa dena. Honela definitzen da:

$$I(w_1; w_2) = \log_2 \frac{P(w_1|w_2)}{Pw_1} \tag{III.22}$$

$I(w_1; w_2)$ eta $I(w_2; w_1)$ balio berekoak dira, eta beraz:

$$\text{PMI}(w_1, w_2) = \log_2 \frac{P(w_1|w_2)}{Pw_1} = \log_2 \frac{P(w_2|w_1)}{Pw_2} = \log_2 \frac{P(w_1, w_2)}{Pw_1 Pw_2} \tag{III.23}$$

Ekuazioaren azken atala da elkarrekiko informazioa kalkulatzeko erabili ohi dena: bi hitzak batera agertzeko probabilitatea zati bi hitzak zori hutsez batera agertzeko probabilitatea (itxarondako probabilitatea).

Kontingentzia-taularen adierazpidean:

$$\text{PMI} = \frac{O_{11}}{E_{11}} = \frac{N \cdot O_{11}}{(O_{11} + O_{12})(O_{11} + O_{21})} \tag{III.24}$$

PMI = 0 denean, bi hitzen artean korrelaziorik ez dagoela ulertu behar da, hitzak independenteak direla. PMI $\gg$ 0 denean, berriz, bi hitzen konbinazioa kolokaziotzat har daiteke. PMI $\ll$ 0 ere izan daiteke, eta horrek esan nahi du konbinazioko hitzek elkar "uxatzeko" joera dutela, hau da, bata agertuz gero, bestea agertzeko probabilitatea zoriak aginduko lukeena baino txikiagoa dela. Horrelako konbinazioei *antikolokazio* deitu ohi zaie (Pearce, 2001); esaterako, *several thanks* eta *powerful tea* antikolokazioak dira, *many thanks* eta *strong tea* baitira ohiko konbinazioak.

Elkarrekiko informazioari maizen egozten zaion desabantaila da maiztasun gutxiko hitzez osatutako konbinazioei ematen dien pisu handia (Smadja et al., 1996). Horrelako hitzen kasuan, itxarondako probabilitatea oso txikia da, eta oso maiztasun txikiko bigramek ere PMI handia izan dezakete. Elkarrekiko informazioa independentziaren neurri egokia da, baina mendekotasuna neurtzeko ez da hain ona, zeren, horretarako, konbinazioa osatzen duten hitzen maiztasunak hartu behar baitira kontuan, eta elkarrekiko informazioak ez du hori egiten (Manning eta Schütze, 1999: 170).

Hori konpondu nahian, maiztasun-atari bat ezarriz agerpen-kopuru minimo batetik beherako kasuak aintzat ez hartzea proposatu izan da (Church

eta Hanks, 1990). Bestetik, MItik eratorritako heuristiko batzuk proposatu dira: MI^2 eta MI^3:

$$\mathrm{MI}^2 = \frac{(O_{11})^2}{E_{11}}; \quad \mathrm{MI}^3 = \frac{(O_{11})^3}{E_{11}} \tag{III.25}$$

Aurreko bi neurriak MI^k formako neurri heuristiko parametriko baten bi kasu jakinak dira (non $k = 2$ eta $k = 3$). Daillek 2-10 bitarteko k-ren balioekin egindako esperimentuetan (Daille, 1994: 139), emaitza onenak $k = 3$ parametroarekin lortu ditu. MI^3 neurria SketchEngine tresnak darabiltzan neurrietako bat ere bada[5].

Maiztasuna (f)

Bi hitz batera agertzeko joeraren neurri oinarrizkoena da maiztasuna. UFak erauzteko neurritzat erabiltzearen motibazioa hau da: UFak konbinazio libreak baino maiztasun handiagokoak dira (Evert, 2005: 89). Evertek AM heuristikoetan sailkatzen du maiztasuna (heuristikoak dira, motibazio teorikoa izan gabe ere, argumentu heuristikoen bidez eratuak direnak edo beste neurri batzuetatik heuristikoki eratorriak direnak).

Hala ere, II.1.2.1 atalean idiosinkrasia estatistikoa definitzean adierazi dugunez, corpus batean hitz-konbinazio bat agertzen den maiztasuna, absolutua zein erlatiboa, ez da, teorikoki behintzat, behar bezain neurri fidagarria konbinazio bat UFa den ala ez ebazteko.

Dena den, enpirikoki egiaztatu behar da maiztasunaz adierazi ditugun zalantzak eta eragozpenak benetan gertatzen diren ala ez. Horregatik, f neurria ere AMen arteko bat izatea komeni da. Sarritan, f oinarri-lerrotzat hartu izan da, a priori sofistikatuagoak diren AMen portaera konparatzeko (Evert, 2005: 117). Aurrerago ikusiko dugunez, hainbat ikertzaile harritu ere egin dira zenbait esperimentutan f neurriak erdietsitako emaitza onez (Krenn eta Evert, 2001; Wermter eta Hahn, 2006).

III.6.4 AMen aplikagarritasuna

Testutik UFak eta, bereziki, kolokazioak erauzteko egin den lan esperimentala oso handia izan da, batez ere azken bi hamarkadetan, eta lan horien emaitzen azterketa konparatiboa eta kritikoa ere egin dute hainbat ikertzailek (Pearce, 2002; Thanopoulos et al., 2002; Evert, 2005; Pecina, 2005;

[5] http://trac.sketchengine.co.uk/raw"-attachment/wiki/SkE/DocsIndex/ske"-stat.pdf

Wermter eta Hahn, 2006). Lan horietako askoren asmoa izan da AM egokiena zein den aurkitzea, ahalik eta xede gehienetarako eta irizpide objektiboetan oinarrituta. Horrek ekarri du ebaluazioa zorrozteko ahaleginak egitea, ikerketen emaitzak modu fidagarrian konparatu ahal izateko.

Gaur egun, azterketa konparatiboak egin dituzten ikertzaile askok aitortzen dute ez dela orotarako neurri idealik, faktore askoren mendekoa baita AMen performantzia: hizkuntza, corpusaren tamaina, testuen erregistroa, erauzi nahi den UFaren egitura sintaktikoa eta idiomatikotasun-maila (esapide idiomatikoak, kolokazioak, euskarri-aditzeko eraikuntzak...) (Evert eta Krenn, 2005b: 452). AM idealik ezean, azken urteetan saio batzuk egin dira AMak konbinatuz emaitzak hobetzeko, batez ere ikasketa automatikoko teknikak aplikatuz (ikus III.10).

III.7 Konposizionaltasun semantikoaren neurketa

Ez-konposizionaltasun semantikoa (edo, kolokazio askoren kasuan, erdikonposizionaltasuna) idiomatikotasunaren ezaugarri nabarmenentzat kontsideratu da (Manning eta Schütze, 1999), edo, II.1.2 atalean azaldu dugun bezala, ezaugarri bakartzat, bi terminoak sinonimotzat jotzea ohiko baita oraindik. Hala ere, UFak erauzteko teknika estandarrak agerkidetzaren neurketa hutsean oinarrituak izan dira duela gutxi arte. Azken hamarkadan, ikerkuntza handia egin da konbinazioen konposizionaltasuna testuetako informazioa prozesatuz automatikoki neurtzeko.

Konposizionaltasunaren printzipioa aurkeztu genuenean, honela adierazi genuen: "osoaren esanahia sintaktikoki konbinatuta dauden osagaien esanahien funtzioa da" (Partee, 1995). Beraz, UF hautagaia den konbinazio bat konposizionala den ala ez jakiteko, osagaien esanahien batura "kalkulatu" egin beharko litzateke nolabait, eta gero konbinazioaren esanahiarekin konparatu (Korkontzelos, 2010: 83). Horretarako, sistema bat behar da hizkuntza-elementuen esanahia modelizatzeko, eta elkarren artean konparatzeko.

Hizkuntza-elementuen esanahia modelizatzeko sortu den teknika ahaltsuena antzekotasun distribuzionalaren hipotesian oinarritzen da. Hipotesia Z. Harrisen ideiei helduta motibatu ohi da (Harris, 1954), eta honelako formulazioak eman zaizkio, besteak beste: "testuinguru berdinetan agertzen diren hitzek antzeko esanahia izaten dute."[6]

[6]Harrisen beraren hitzetan (Harris, 1970: 786): "Difference of meaning correlates with difference of distribution." Ideia hori erlazionatuta dago, nolabait, kolokazioaren kontzeptua proposatu zuen J.R. Firthen aipu ezagun honekin: "You shall know a word by the company it keeps" (Firth, 1957).

Aipatu hipotesia betetzen ez den kasuak badaudela aitortu da (Korkontzelos, 2010: 76), hala nola hiperonimo-hiponimo erlazioa duten hitz-bikoteen kasua; edo hitz bietako bat polisemikoa izanik, beste hitzaren adiera hitz polisemikoaren adiera nagusia (maiztasun handienekoa) ez izatea.

Nolanahi ere den, hizkuntzaren prozesamendu automatikoan, antzekotasun distribuzionala oso teknika erabilia da hitzen ezaugarri semantikoak karakterizatzeko, batez ere hitzen adiera-desanbiguazioan edo HADean (*word-sense dissambiguation*, WSD) (Agirre eta Edmonds, 2006). Fraseologia konputazionalaren arloan ere aplikatu da, hitz-konbinazioen konposizionaltasuna neurtzeko edo kuantifikatzeko, eta horretarako ideia gakoa Linek (1999) eman zuen:

> ≪The intuitive idea behind the method is that the metaphorical usage of a noncompositional expression causes it to have a different distributional characteristic than expressions that are similar to its literal meaning.≫

Horrenbestez, honela formula dezakegu antzekotasun distribuzionalaren hipotesia, UFak karakteritzatzeko atazara egokituta:

> UFaren testuingurua eta haren osagaien testuinguruak zenbat eta desberdinagoak izan, UFa ez-konposizionala izateko aukera handiagoa.

Hipotesi hori egiaztatzeko, beraz, hitz-konbinazioen testuinguruak beren osagaien testuinguruekin konparatu behar dira. Lehen "esanahia modelizatu beharra" aipatu dugu, eta orain, esanahia testuinguruaren bidez errepresenta daitekeelako hipotesian oinarrituta, esan dezakegu testuingurua dela modelizatu behar duguna. Hori egindakoan, konparatu nahi ditugun testuinguruen errepresentazioen arteko antza kalkulatu behar da. Bi egiteko horietan barneratuko gara hurrengo bi ataletan.

III.7.1 Testuinguruaren errepresentazioa (modelizazioa)

Testuingurua nola errepresentatu lantzen hasi aurretik, testuinguruaren beraren definizioa zehaztu beharra dago. Hurrengo paragrafoetan, erabakigai edo parametro nagusiak azalduko ditugu (Weeds, 2003: 18-21; Korkontzelos, 2010: 71-75). Aplikazioaren eta hizkuntzaren arabera, aldatu egin daiteke parametro horietako aukera egokiena zein den, eta egiaztapen esperimentala izaten da modurik ohikoena hori ebazteko. Dena den, badira, gure ikergaia

UFen konposizionaltasuna neurtzea dela kontuan izanik, aurrez egin litezkeen iruzkin batzuk.

- Testuinguruaren helmena: dokumentua, paragrafoa, perpausa, aurreko eta ondorengo testu-hitz kopuru jakin bat...

  A priori, badirudi hitz edo hitz-konbinazio baten gertuko testuingurua dela semantikoki informatiboena (Lin, 1997; Schütze eta Pedersen, 1997), eta konbinazioa agertzen den dokumentu osoa kontuan hartzea errelebantea izango ez den informazioa erabiltzea dela. Horregatik, testuingurua ez da hain zabal izaten, eta gehienetan perpausa edo hitzaren inguruko leiho jakin baten barneko hitzak erabili ohi dira.

- Testuinguruko hizkuntza-elementu bat errepresentazioan sartzeko, zer erlazio izan behar duen errepresentatu nahi dugun hitzarekiko: agerkidetza, erlazio sintaktikoa...

  Agerkidetzan oinarritutako sistemari *bag-of-words* eredua deitu ohi zaio. Hitz baten testuinguruaren helmenaren barnean agertzen diren beste hitz guztiak sartzen dira errepresentazioan, zein ere den hitzen arteko mendekotasun sintaktikoa. Beste aukera, berriz, halako mendekotasun-erlazioa duten hitzak kontuan hartzea da. *Bag-of-words* ereduaren oinarrizko asuntzioa da hitz baten inguruan dauden hitz guztiak direla haren esanahia errepresentatzeko errelebanteak, eta horrek ereduaren inplementazioa erraza izatea dakar berekin; baina aski ebidentzia bide dago eredu hori sinplifikazio bat dela, eta perpaus barneko zein perpausen arteko erlazio sintaktikoak garrantzitsuak direla (Padó eta Lapata, 2007: 2). Aditzen esanahia konparatzean, esaterako, hartzen dituzten objektuen multzoak konpara daitezke, zehazki, haien testuinguruan diren izen guztien multzoak konparatu beharrean.

  Bi teknika horien inguruko eztabaida bizirik dago, eta, gainera, aukera bataren edo bestearen egokitasuna aplikazioaren araberakoa ere izan daiteke. Oro har, esan liteke (Weeds, 2003: 20-21) *bag-of-words* ereduak erakutsi duela IR eta kideko arloetan eraginkorra dela; bestalde, erlazio sintaktikoak kontuan hartzen dituzten errepresentazioak garatzeko saio berri gehienak lexiko-semantikan egin dira. Ikusiko dugunez, hitz-konbinazioen konposizionaltasuna neurtzeko saiakuntza gehienek *bag-of-words* eredua aplikatu dute (Korkontzelos, 2010: 78).

- Testuinguruko hizkuntza-elementuen zer alderdi eramango den errepresentaziora: hitz-forma, lema, kasua, mugatasuna...

Lehen ideia intuitiboa da testuko hitzen semantikari ekarpen handiena hitzaren lemak egiten diola, hori dela esanahiaren muina, eta hitz-forma (*token*), berriz, lemaren forma flexionatua dena, ez dela hain informazio errelebantea. Esaterako, testuinguru hauek izanik,

(11) Autoa abiadura handiegian zihoan, eta bidetik atera zen

(12) Errepidea bustita zegoen, eta abiadura handiegia zeramaten hainbat autok talka egin zuten

ez dirudi lehen adibideko *bide* eta bigarreneko *errepide* hitzen esanahiei ekarpenik egiten dienik batean *autoa* eta bestean *autok* formak agertzeak. Hori gabe, bi esaldietan ageri den *auto* lema baino ez da esanguratsua. Beraz, lemak kontuan hartuta, testuinguruen errepresentazioak murritzagoak dira (elementu desberdin gutxiago daude), bereziki hizkuntza eranskarietan. Horrek datu-barreiatzea murrizten du, estatistika esanguratsuagoak izan daitezke, eta, ataza batzuetan behintzat, emaitza hobeak lortu dira (Sahlgren, 2006: 76, Lopez de Lacalle, 2009: 32).

- Errepresentazioan kategoria guztietako hitzak sartuko diren ala eduki-hitzak (*content-bearing words*) bakarrik, eta, kasu horretan, eduki-hitzen sailean zer kategoria sartzen den.

  Semantikaren ikuspegitik, iritzi nagusia da funtzio-hitzak direla (erakusleak, zenbatzaileak, izenordainak, juntagailuak...) informazio bereizgarri gutxien eskaintzen dutenak (Sahlgren, 2006: 38; Korkontzelos, 2010: 72). Eduki-hitzen artean, eskuarki izenak, aditzak eta adjektiboak kontsideratu ohi dira, eta, horietatik, izenak bide dira informatiboenak (Agirre et al., 2006). Horregatik, ohikoa da testuinguruaren errepresentazioan funtzio-hitzak ez sartzea, batez ere *bag-of-words* ereduan, zeinetan hori baita agerkidetza hutsa erabiltzearen ondorioz sortzen den zarata murrizteko modu bakarra.

- Errepresentazioan sartuko den hizkuntza-elementua zer informazioz hornitu edo nola karakterizatuko den.

  Eredu batzuetan, hitz baten testuinguruetan beste hitz bat agertu den ala ez baino ez da zehazten ($0-1$); beste batzuetan berriz, testuinguru-hitz bakoitzari datu estatistiko bat gehitzen zaio: agerkidetza-maiztasun absolutua edo erlatiboa, *tf-idf* balioa, edo III.6.3 atalean azaldu ditugun AMetako bat (Korkontzelos, 2010: 81).

Hitzen testuingurua errepresentatzeko erabiltzen den ohiko eredua IR arlorako garatu zen *Vector Space Model* (VSM) edo *bektore-espazioaren eredua* izeneko eredu ezagunaren egokitzapena da, eta *Word Space Model* (WSM) edo *hitz-espazioaren eredua* izendapena ere eman ohi zaio (Schütze, 1993; Sahlgren, 2006). VSMren sorrera Salton et al.-en (1975) lanari aitortzen zaio. Gerora, beste zenbait arlo eta atazatara aplikatu edo egokitu da (Weeds, 2003: 22-35); esaterako, adiera-desanbiguazioa (Agirre et al., 2006), lexiko-eskurapena (Curran eta Moens, 2002), thesaurus-eraketa eta ontologia-ikasketa (Cimiano, 2006) edo corpus konparagarrietatik terminologia elebiduna erauztea (Fung eta Yee, 1998; Saralegi et al., 2008). Laster ikusiko dugunez, VSM hitz-konbinazioen konposizionaltasuna neurtzeko ere aplikatu da.

Jatorriz, IRko VSMn "termino-dokumentu" bektoreak edo matrizeak eratzen dira, eta WSMn, berriz, "hitz-testuinguru" (*word-context*) erakoak (Turney et al., 2010: 146). Hitz baten testuinguruaren errepresentazioa egiteko, bektore bat eratzen da, non lehen zutabean testuinguruko hitzak dauden, eta, bigarren zutabean, hitz horietako bakoitzaren ezaugarri bat edo "pisu" bat, hala nola f, baldintzazko probabilitatea, *tf-idf*, PMI...

abiadura	2
atera	1
bide	1
busti	1
egin	1
errepide	1
handi	2
talka	1

III.2 Taula: (11) eta (12) adibideetako *auto* hitzaren bektorea.

Esaterako, III.2 taulan, (11) eta (12) adibideetako *auto* hitzaren bektorea dugu, *bag-of-words* ereduaren eta eduki-hitzen lemak kontuan hartzeko irizpidearen arabera, helmentzat esaldia eta pisutzat maiztasun absolutua erabiliz (aditz laguntzailea funtzio-hiztzat hartuta).

Bektore horrek 8 dimentsio ditu. Corpus bateko hitzen bektoreak sortzean, hitz guztien bektoreak $n \times n$ matrize batean errepresenta ditzakegu (n lema- edo hitz-kopurua izanik) (Lund eta Burgess, 1996). Gure corpusa aurreko bi adibideak direla pentsatuz, III.3 taulako matrizea lortuko genuke.

Adibidez, *bide*, *errepide* eta *talka* hitzen esanahiak zenbateraino diren antzekoak neurtzeko, bektore hauek konparatu behar ditugu:

bide	(1, 1, 1, 0, 0, 0, 0, 1, 0)
errepide	(1, 0, 1, 0, 1, 1, 0, 1, 1)
talka	(1, 0, 1, 0, 1, 1, 1, 1, 0)

Hitza	**Agerkidea**								
	abiadura	atera	auto	bide	busti	egin	errepide	handi	talka
abiadura	0	1	2	1	1	1	1	2	1
atera	1	0	1	1	0	0	0	1	0
auto	2	1	0	1	1	1	1	2	1
bide	1	1	1	0	0	0	0	1	0
busti	1	0	1	0	0	1	1	1	1
egin	1	0	1	0	1	1	1	1	1
errepide	1	0	1	0	1	1	0	1	1
handi	2	1	2	1	1	1	1	0	1
talka	1	0	1	0	1	1	1	1	0

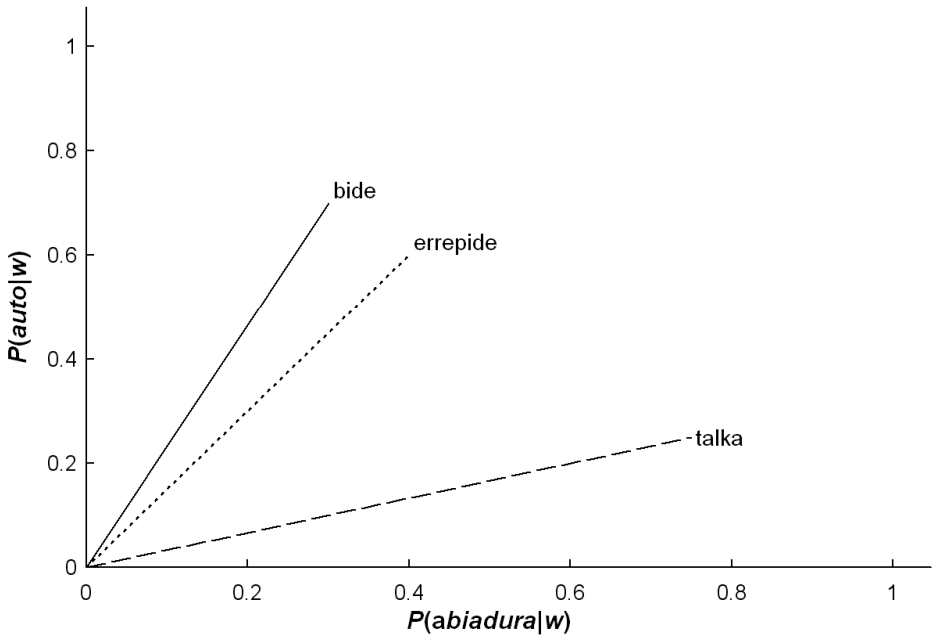
III.3 Taula: (11) eta (12) adibideetatik ateratako agerkidetza-taula.

Formalki, bektorea bektore-espazio bateko elementua da, n osagai edo koordenatu dituena (Sahlgren, 2006: 27). Aurreko adibidean, 9 dimentsioko bektoreak ditugu. Grafikoki irudikatu ahal izateko, pentsa dezagun 2 dimentsio baino ez ditugula (esaterako, *abiadura* eta *auto*), eta corpus handiago bateko esaldiak prozesatu ondoren, III.4 taulako bektoreak atera ditugula. Eskuinean diren baldintzazko probabilitateak kalkulatzeko, sinplifikatzearren, pentsatu dugu taulako horiek berak direla konparatu nahi ditugun hitzen agerkidetza bakarrak.

Hitza	**Agerkidea**			$P(c\|w)$	
	abiadura	auto	totala		
bide	78	182	260	0,30	0,70
errepide	44	66	110	0,40	0,60
talka	45	15	60	0,75	0,25

III.4 Taula: *bide*, *errepide* eta *talka* hitzen *abiadura* eta *auto* hitzekiko agerkidetzak.

Bektoreak 2 dimentsioko grafiko batean irudikatu dira III.1 irudian. Bi bektore bata bestetik zenbat eta gertuago egon, hainbat handiagoa da haien arteko antzekotasun distribuzionala.



III.1 Irudia: abiadura – auto espazioan, *bide*, *errepide* eta *talka* hitzen bektoreak.

Hitz-konbinazioen konposizionaltasuna neurtu nahi dugu eta, hitz bakunekin egin dugun bezala, hitz-konbinazio baten testuingurua bektore baten bidez errepresenta dezakegu. Orain alderdi giltzarria da hori zerekin konparatu. Bi aukera ditugu:

- Konbinazioaren esanahi konposizionala "eraikitzea", hitz bakunen bektoreak nolabait konbinatuz.

  Lehenago esan dugu osagaien esanahien batura "kalkulatu" egin beharko litzatekeela nolabait, eta gero konbinazioaren esanahiarekin konparatu. AMen kasuan bezala, non konbinazioaren itxarondako maiztasun edo probabilitate bat estimatu behar baitzen, orain ere konbinazioaren esanahi konposizionala "aurresan" egin behar dugu. Osagaien bektoreak erabiliz konbinazioaren bektore konposizionala modelizatzeko, hiru teknika proposatu dira (Guevara, 2010: 34): batuketa

($v1_i + v2_i = v3_i$), biderketa puntuala ($v1_i \times v2_i = v3_i$) eta tentsore-biderketa ($v1 \oplus v2 = v3$).

- Konbinazioaren bektorea osagai bakoitzaren bektorearekin konparatzea, eta gero bi konparazio horien emaitzak nolabait konbinatzea.

  Horrela jokatzeak badu abantaila bat: jakina da kolokazio batzuk erdikonposizionalak direla, eta, gure ikergaia den `izena+aditza` konbinazioen kasuan, aditzaren semantika ohikoa ez izateari zor zaiola hori, ezen ez izenari. Beraz, interesgarria da hipotesi hori enpirikoki egiaztatzea, eta horretarako, izenaren eta aditzaren semantika bereiz konparatu behar dira konbinazioaren semantikarekin (Wulff, 2008).

Bietan ere, garrantzitsua da osagaien bektoreak eratzeko zer testuinguru erabiltzen diren. Izan ere, aski intuitiboa da pentsatzea hobe dela osagai bakunen testuinguruetan ez sartzea konbinazioaren testuinguruak (Katz eta Giesbrecht, 2006; Garrão et al., 2006). Horren arabera, *adarra jo* konbinazioaren konposizionaltasuna neurtzean, *adar* eta *jo* osagaien testuinguruetan ez genituzke sartuko *adarra jo* konbinazioaren testuinguruak. Horrela jokatuz, erabat bereizten dugu konbinazioaren eragina osagaien testuinguruen modelizazioan, eta konparazioa adierazgarriagoa izan daiteke.

III.3 taulako bezalako agerkidetza-matrizeak tamaina handiko corpusekin eratzen direnean, dimentsionaltasun handiaren arazoa sortzen da. Batetik, hitz-espazioaren ereduak datu-kopuru handia behar du distribuzioaren ebidentzia estatistikoa lortzeko, baina dimentsio-kopuru erraldoiak zaildu egin dezake horren prozesamendua, konputazio-eskakizuna handiegia izan baitaiteke (Sahlgren, 2006: 37). Bestetik, datu-barreiatzearen arazoa ere gertatzen da, hau da, agerkidetza-matrizeko gelaxka askoren balioa zero da.

Arazo horiek konpontzeko, dimentsionaltasuna gutxitzea da landu nahi izan den alderdia. Lehen aukera da eduki-hitzez mintzatu garenean aurkeztu duguna, hau da, corpuseko hitz guztiak ez errepresentatzea matrizean, eta semantikari ekarpena egiten diotenak baino ez erabiltzea. Hala ere, funtzio-hitzak kenduta ere, lexikoaren parte handiena geratzen da. Beste aukera bat izan daiteke agerkidetzaren aldetik esanguratsuak ez diren hitzak iragaztea, hau da, oso maiztasun handiko edo txikiko hitzak. Jakina da, Zipfen legearen arabera, maiztasun handiko hitzen proportzioa oso txikia dela, eta, beraz, horiek kontuan ez hartzea ez da oso eraginkorra dimentsionaltasuna gutxitzeko. Eztabaidagarriagoa da, berriz, maiztasun apaleko hitzak baztertzearen eragina (Weeds, 2003: 173-178; Sahlgren, 2006: 104).

Dimentsionaltasuna gutxitzeko garatu den teknika ezagunena *Singular Value Decomposition* (SVD) edo *balio singularretan deskonposatzea* da. Tek-

nika hori erabiltzen du bektore-espazioaren inplementazio hedatu ezagunenetakoak, *Latent Semantic Analysis* (LSA) edo *ezkutuko semantikaren analisia* izenekoak (Deerwester et al., 1990; Dumais, 2004; Zelaia et al., 2011).

SVDren oinarri matematikoen azalpen zehatza Berry et al. (1995), Lopez de Lacalle (2009), eta Zelaia et al.-en (2011) lanetan aurki daiteke. Hemen aski da azaltzea SVD matrizeen faktorizazio-teknika bat dela, jatorrizko matrizea hiru matrize txikiagotan deskonposatzen duena. Matrize horiek jatorrizko matrizearen faktore linealki independenteak dituzte (balio singularrak). Matrizeak biderkatzean balio singular txikienei ezikusia eginez, emaitza jatorrizko matrizearen hurbilpena da. Horri *truncated SVD* deritzo, *bektore-espazio murriztua* edo *espazio semantikoa* (Zelaia et al., 2011: 93).

SVD informazio-berreskuratzean erabiltzeko arrazoiak agerikoak dira: antzeko agerkidetza-patroiak dituzten hitzak multzokatuz, sistema gai izan daiteke kontsultan ez dauden hitzen sinonimoak edo kuasisinonimak dituzten dokumentuak ere berreskuratzeko. Baina WSM eredurako, zein izan daiteke onura? SVDren abantailez diharduela, Lopez de Lacallek (2009) *high-order co-occurrence* ("goi-mailako agerkidetza") eta *latent meaning* ("esanahi sorra"' edo "ezkutukoa") azaltzen ditu. SVDri esker, azalekoak edo esplizituak ez diren zeharkako erlazioak atzeman daitezke, eta horrek antzekotasunaren neurketa hobetzeko aukera ematen bide du. Horiek dira, hain zuzen ere, aurreko galderari erantzuteko, Sahlgrenek (2006) aipatzen dituenak.

III.7.2 Antzekotasun distribuzionaleko neurriak

Testuinguruen errepresentazioa diren bektoreak konparatzeko, antzekotasun distribuzionaleko neurriak erabiltzen dira (*distributional similarity measures*). n dimentsioko $\vec{x}$ eta $\vec{y}$ bi bektore izanik, eta $x(i)$ eta $y(i)$ bektoreen i-garren dimentsioko balioak direla, ondoko taulan ageri den bezala konputatzen dira literaturan aipatuenak diren antzekotasun-neurriak (Weeds, 2003: 47-59; Korkontzelos, 2010: 77-81).

Lehen hiru neurriek geometrian dute jatorria, Jaccard eta Dice koefizienteek konbinatorian, eta gainerakoak ordezkagarritasunean oinarritutako neurriak dira (Weeds, 2003: 47-59).

Aurreko neurrien artean, kosinua da erabiliena. VSM ohikoetan ez ezik, LSAren inplementazio gehienetan ere erabiltzen da. Bektore-espazioan bi bektorek osatzen duten angeluaren kosinua da, eta $[0-1]$ arteko balioa izan dezake; zenbat eta kosinu handiagoa, hainbat handiagoa bi hitzen arteko antzekotasun distribuzionala.

Manhattan distantzia (L_1 norm)

$$L_1(\vec{x},\vec{y}) = \sum_{i\in[1,n]} |x(i) - y(i)| \tag{III.26}$$

Distantzia euklidearra (L_2 norm)

$$L_2(\vec{x},\vec{y}) = \sqrt{\sum_{i\in[1,n]} (x(i) - y(i))^2} \tag{III.27}$$

Kosinua

$$cos(\vec{x},\vec{y}) = \frac{\sum_{i\in[1,n]} x(i) \times y(i)}{\sqrt{\sum_{i\in[1,n]} x(i)^2 \times \sum_{i\in[1,n]} x(i)^2}} \tag{III.28}$$

Jaccard koefizientea

$$J(\vec{x},\vec{y}) = \frac{|\{i \in [1,k] : x(i) \neq 0 \wedge y(i) \neq 0\}|}{|\{i \in [1,k] : x(i) \neq 0 \vee y(i) \neq 0\}|} \tag{III.29}$$

Dice koefizientea

$$dice(\vec{x},\vec{y}) = \frac{2 * |\{i \in [1,k] : x(i) \neq 0 \wedge y(i) \neq 0\}|}{|\{i \in [1,k] : x(i) \neq 0\}| + |\{i \in [1,k] : y(i) \neq 0\}|} \tag{III.30}$$

Kullback-Leibler dibergentzia

$$D_{KL}(\vec{x} \parallel \vec{y}) = \sum_{i\in[1,n]} x(i) \log \frac{x(i)}{y(i)} \tag{III.31}$$

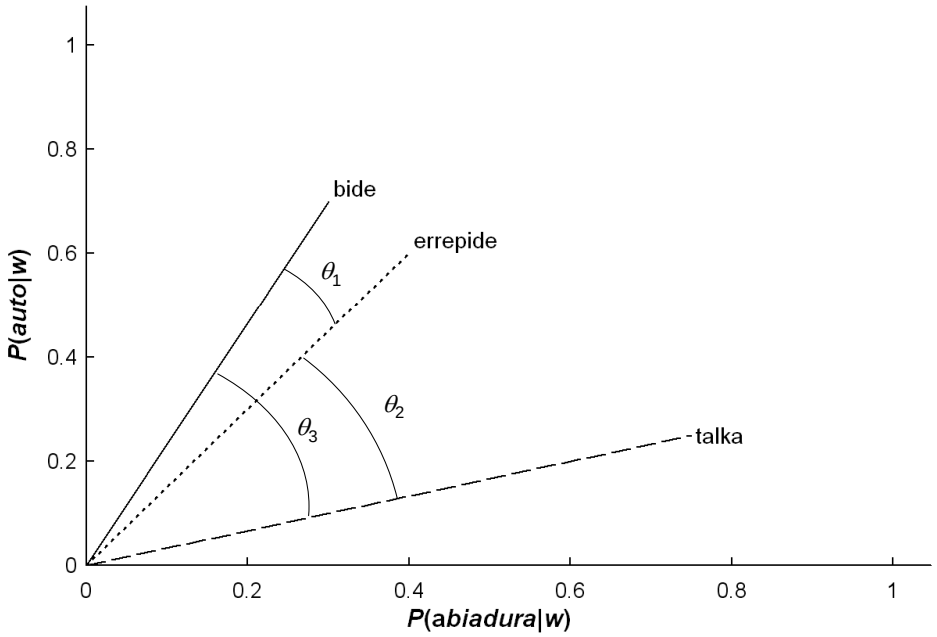
Jensen-Shannon dibergentzia

$$JSD(\vec{x},\vec{y}) = \frac{1}{2}[KL(\vec{x} \parallel \frac{1}{2}(\vec{x}+\vec{y})) + KL(\vec{y} \parallel \frac{1}{2}(\vec{x}+\vec{y}))] \tag{III.32}$$

α-skew dibergentzia

$$s_\alpha(\vec{x},\vec{y}) = KL(\vec{x} \parallel \alpha \cdot \vec{x} + (1-\alpha) \cdot \vec{y}) \tag{III.33}$$

III.2 grafikoan ikus dezakegu *bide* eta *errepide* hitzen arteko antza handiagoa dela, haietako edozeinek *talka*rekin duen antza baino.



III.2 Irudia: abiadura – auto espazioan, *bide*, *errepide* eta *talka* hitzen bektoreen arteko angeluak.

Kosinuaren balioak:

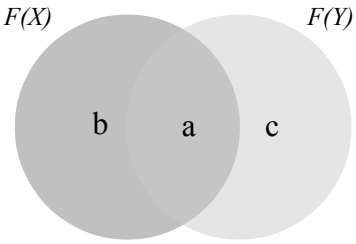
$$\cos(\textit{bide}, \textit{errepide}) = \cos\theta_1 = 0,98$$
$$\cos(\textit{errepide}, \textit{talka}) = \cos\theta_2 = 0,78$$
$$\cos(\textit{bide}, \textit{talka}) = \cos\theta_3 = 0,66$$

Bestetik, Jaccard eta Dice koefizienteak neurri konbinatorioak dira, eta konparagaiek, hau da, gure bektoreek, partekatzen dituzten agerkideen kontaketan oinarrituta daude. $F(X)$ eta $F(Y)$ bi bektoreen balioen multzoak badira, honela adierazten da Jaccard koefizientea, III.3 irudiko multzo-ebakidurak erabiliz:

$$J(X,Y) = \frac{|F(X) \cap F(Y)|}{|F(X) \cup F(Y)|} = \frac{a}{a+b+c} \tag{III.34}$$

Dice koefizientea, berriz:

$$D(X,Y) = \frac{2 \times |F(X) \cap F(Y)|}{|F(X)| + |F(Y)|} = \frac{2a}{2a+b+c} \tag{III.35}$$



III.3 Irudia: $F(X)$ eta $F(Y)$ bektoreen balioen multzoak, eta multzo horien ebakidurak.

Jaccard eta Dice koefizienteek ranking berdinak sortzen dituzte, bien arteko bihurketa monotonikoa baita: $J = D/(2 - D)$ (Evert, 2005: 56).

Kullback-Leibler dibergentziak (KL dibergentziak), edo entropia erlatiboak, bi banaketaren arteko distantzia neurtzen du (Kullback eta Leibler, 1951, Weeds, 2003: 47-59). Zehazki, $D_{KL}(P \parallel Q)$ dibergentziak neurtzen du zer informazio-galera gertatzen den Q banaketa erabiltzen denean P banaketaren hurbilketatzat. Neurri asimetrikoa da, hau da: $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$. Bestetik, definitua izateko, bete behar da $P(i) > 0$ denean $Q(i) > 0$ ere izatea, bestela $D_{KL}(P \parallel Q) = \infty$. Ondorioz, komenigarria da *smoothing* tekniken bidez ez-zero probabilitateak esleitzea behatu gabeko gertakariei (Dagan, 2000: 470). Horrek konplika dezake inplementazioa, eta, eragozpen horiek saihesteko, Jensen-Shannon dibergentzia erabil daiteke, neurri simetriko bat egiteko asmoz proposatutako egokitzapena dena (Lee, 1999: 26). Beste aukera bat α-skew dibergentzia da (Weeds, 2003: 55-56).

III.7.3 Antzekotasun distribuzionalaren neurketa UFen konposizionaltasuna karakterizatzeko

UFen konposizionaltasuna neurtzeko lehen ikerlanen artean, ekarpen nabaria da Berry-Roggherena (1974). *R-value* izeneko neurri bat proposatzen du *verb-particle constructions* (`VPC`) direlakoen konposizionaltasuna neurtzeko (esaterako, *live in, believe in...*). Neurri hau kolokazio-erauzketaren arloan garatu da, eta normalean ez da antzekotasun distribuzionaleko ohiko neurrien artean aipatzen.

R delako balioa bi kopuru hauen arteko zatidura da: `VPC`ak eta partikulak partekatzen dituzten kolokatuen kopurua, eta `VPC`aren beraren kolokatuen kopurua. VPC eta P, hurrenez hurren, konbinazioaren kolokatuen eta partikularen kolokatuen multzoak badira:

$$R = \frac{|VPC \cap P|}{|VPC|} \tag{III.36}$$

Lehen Jaccard koefizientea aurkeztu dugunean erabili dugun adierazpidean:

$$R = \frac{a}{a+b} \tag{III.37}$$

R balioa $[0-1]$ tartean alda daiteke, 0 denean `VPC`a erabat ez-konposizionala da, eta 1 denean, berriz, erabat konposizionala. Berry-Roghek kolokatu esanguratsuak soilik hartzen ditu kontuan, eta, horiek hautatzeko, kolokatuen z neurriaren (*z-score*) minimo bat hartzen du.

Berry-Rogheren proposamena hobetu nahian, Wulffek (2008) R balioaren bi hedapen proposatu ditu. `V-NP` (*verb+noun-phrase*) konbinazioen konposizionaltasuna neurtzeko, osagai bakoitzarekiko R balioak neurri bakarrean konbinatzen ditu.

$$\text{comp}_{\text{V-NP}} = \text{contribution}_{\text{V}} + \text{contribution}_{\text{NP}} \tag{III.38}$$

Osagaien ekarpena bakoitzaren R balio haztatua da, hau da, R balioa bider osagaiaren ekarpenaren pisua. R balioak hauek dira:

$$R_{\text{V}} = \frac{\text{n colls}_{\text{pattern}} \text{ in colls}_{\text{V}}}{\text{n colls}_{\text{pattern}}} = \frac{a}{a+b} \tag{III.39}$$

$$R_{\text{NP}} = \frac{\text{n colls}_{\text{pattern}} \text{ in colls}_{\text{NP}}}{\text{n colls}_{\text{pattern}}} = \frac{e}{e+f} \tag{III.40}$$

Bestetik, R balioak haztatzeko, ideia nagusia da osagai batek zenbat eta kolokatu gehiago izan, hainbat eta pisu handiagoa izan behar lukeela bere R balioak. Ideia hori bi eratara inplementatu du Wulffek.

Lehen hedapenean, osagai baten kolokatuek bi osagaien kolokatuen bilduran duen pisua hartzen da kontuan:

$$\text{weight}R_{\text{V}} = \frac{\text{n colls}_{\text{V}}}{\text{n colls}_{\text{V}} + \text{n colls}_{\text{NP}}} \tag{III.41}$$

$$\text{weight}R_{\text{NP}} = \frac{\text{n colls}_{\text{NP}}}{\text{n colls}_{\text{V}} + \text{n colls}_{\text{NP}}} \tag{III.42}$$

Jaccard koefizientea aurkeztu dugunean erabili dugun adierazpidean:

$$\text{comp}_{\text{V-NP}} = \left(\frac{a+c}{(a+c)(e+g)} \times \frac{a}{a+b}\right) + \left(\frac{e+g}{(a+c)(e+g)} \times \frac{e}{e+f}\right) \quad \text{(III.43)}$$

Bigarren hedapenean, konbinazioak eta osagaiak partekatzen dituzten kolokatuen eta osagaiaren kolokatuen arteko arrazoia erabiltzen da:

$$\text{share}R_{\text{V}} = \frac{\text{n colls}_{\text{pattern}} \text{ in colls}_{\text{V}}}{\text{n colls}_{\text{V}}} \quad \text{(III.44)}$$

$$\text{share}R_{\text{NP}} = \frac{\text{n colls}_{\text{pattern}} \text{ in colls}_{\text{NP}}}{\text{n colls}_{\text{NP}}} \quad \text{(III.45)}$$

$$\text{comp}_{\text{V-NP}} = \left(\frac{a}{a+c} \times \frac{a}{a+b}\right) + \left(\frac{e}{e+g} \times \frac{e}{e+f}\right) \quad \text{(III.46)}$$

Horrez gain, Wulffek kolokatu esanguratsuen ehuneko desberdinak hartzen ditu kontuan, eta Fisherren test zehatza erabiltzen du kolokatu esanguratsuak hautatzeko.

Linen (1999) lana ere aitzindaritzat jo ohi da konposizionaltasunaren neurketan. Hala ere, zenbait autorek ohartarazi dutenez, osagaien ordezkagarritasunaren neurketan dago oinarrituta, eta ez du, *stricto sensu*, konposizionaltasuna neurtzen (Baldwin et al., 2003):

> ≪We claim that substitution-based tests are useful in demarcating MWEs from productive word combinations, but not in distinguishing the different classes of decomposability. Simple decomposable MWEs such as *motor car* fail the substitution test not because of nondecomposability, but because the expression is institutionalised to the point of blocking alternates.≫

Horregatik, guk malgutasun lexikalaren atalean sailkatu dugu Linen lana (III.9 atala).

Lan aipagarrienen artean, Garrão et al.-ek (2006) portugesezko `V+NP` konbinazioekin egindako ikerkuntza dugu. Bektoreak osatzeko, paragrafo osoa hartzen dute kontuan, eta, bektoreen arteko antzekotasun-neurritzat, kosinua erabiltzen dute. Ebaluazio kualitatiboa egiten dute, eta emaitzek gehienetan konposizionaltasuni buruzko intuizioak berresten dituztela diote. Fazly eta Stevenson (2007) ere kosinuaz baliatzen dira ingelesezko `VN`

konbinazioen konposizionaltasuna neurtzeko, baina ±15 leiho baten barneko izenak kontuan hartuz bakarrik. UFen beste propietate batzuen neurketekin konparatua, emaitzak onak ez direla nabarmentzen dute.

Distributional Semantics and Compositionality (DiSCo) izeneko ataza partekatuan (Biemann eta Giesbrecht, 2011), aurkeztutako zenbait sistema bektore-espazioan oinarritu dira: Reddy et al. (2011) –emaitza onenak lortutakoa–, Maldonado-Guerra eta Emms (2011) eta Garrido eta Peñas (2011). Ebaluaziorako, eskuz antolatutako erreferentzia landu dute, Amazon Turk sistema erabiliz (Biemann eta Nygaard, 2010). UF hautagai bakoitzari 0-10 bitarteko konposizionaltasun-maila esleitu behar diote atazan izena ematen duten pertsonek; neurritzat, Spearman ρ eta Kendall τ erabiltzen dituzte.

LSA ere erabili da zenbait ikerlanetan. Schone eta Jurafskyk (2001) probatu zuten aurrenekoz, ingelesezko UFak erauzteko hainbat agerkidetza-teknikaren emaitzak aztertu ondoren, horiek hobetzeko asmoz. Ebaluaziorako, WordNet eta Interneteko hiztegi-baliabide libreak erabili dituzte. Haien hitzetan, emaitza etsipengarria izan zen: AMekiko hobekuntzarik ez lortzea ez ezik, emaitzak agerkidetzakoak baino okerragoak izan ziren.

Baldwin et al.-en (2003) lanean ere aplikatu da LSA, baina ez UFak erauzteko, egiaztatutako UFak konposizionaltasunaren arabera sailkatzeko baizik. Haien iritziz, aurreko egileen emaitza eskasen arrazoia LSA erauzketarako erabili izana da. `NN` (`noun+noun`) eta `VPC` (`verb+particle`) egiturako unitateekin egiten dute lan, eta, ebaluaziorako, WordNet erabiltzen. Emaitzek ez dute korrelazio handirik WordNeten oinarritutako anztekotasunekin. Dena den, hiponimia-erlazioak kontuan hartzen direnean (bigramaren sintagmaren burua bigramaren hiperonimoa izatea), korrelazioa hobetu egiten da, bereziki `NN` konbinazioen kasuan.

Katz eta Giesbrechtek (2006) Infomap softwarea[7] erabili dute alemanezko `preposition+noun+verb` konbinazioak LSA bidez sailkatzeko. Ebaluaziorako, Krennen (2000) erreferentzia erabili dute; F neurriaren emaitza onenak kosinuaren ataria 0,1-0,2 artean kokatuta lortu dituzte.

Berriki, Krcmár et al.-ek (2013) kritikatu dute Baldwin et al.-en (2003) lanean LSA eredua sortzean hitz-konbinazioak hitz bakun gisa tratatzea, zeren horrek aldarazten baitu konbinazioaren osagaiek testuinguruak. Hori saihesteko, konbinazioen bektoreak gehitu egin dituzte, halako moldez non hitzen bektoreak atxikitzen baitira.

Beste era bateko metodoa adiera-indukzioaren bidezkoa da (*sense induction*). Grafoetan oinarritutako adiera-indukzioa erabili dute Korkontzelos

[7] `http://infomap-nlp.sourceforge.net/`

eta Manandharr ek (2009), ingelesezko `noun+noun` eta `adjetive+noun` konbinazioak konposizionaltasunaren arabera sailkatzeko. Horretarako, konbinaziorako induzitutako adierak eta haren buruarenak konparatzen dituzte, Jaccard indizearen eta Jensen-Shannon dibergentziaren bidez. Ebaluazio-erreferentzia WordNetetik eskuratutako 6 287 hautagaitatik eratu dute, 19 konbinazio ez-konposizional eta beste hainbeste konposizional erabiliz. Oinarri-lerrotzat SemEval-2007 atazako emaitzak erabiliz, Agirre eta Soroaren (2007) zehaztasunaren 5 puntuko hobekuntza aldarrikatzen dute.

III.8 Malgutasun morfosintaktikoaren neurketa

Fraseologiak betidanik UFen finkotasunaren propietateari garrantzi handia eman badio ere, finkotasun morfosintaktikoa testuetatik automatikoki neurtzeko ikerkuntza-lana ez da aurreko bi propietateen kasuan bezain ugaria. Lan gehienak 2000ko hamarkadan hasi ziren argitaratzen, ingeleserako batez ere.

Finkotasun morfosintaktikoaren propietatea, eskuarki, malgutasun erlatiboaren bidez neurtu ohi da, hau da, konbinazioaren portaeraren eta erreferentzia-portaera batekin konparatuz. UF hautagai baten portaera erreferentzia-portaeratik zenbat eta urrunago, hainbat eta zantzu handiagoak hautagaia esapide idiomatikoa edo kolokazioa izateko. Oinarrizko ideia hori metodologia zehatz batean inplementatzeko, hiru alderdi hauek zehaztu behar dira:

- Portaera morfosintaktikoa definitzean kontuan hartuko diren fenomenoak edo "aldakuntza morfosintaktikoak". Alderdi hau aztergaitzat ditugun konbinazioen egitura morfosintaktikoaren eta hizkuntzaren araberakoa da.

- Erreferentzia-portaera nola definitzen den. Literaturan bi aukera hauek erabili dira:

  - Portaera orokorra: kategoria-osaera bereko konbinazioen batez besteko portaera.

  - Osagaien portaera: konbinazioaren osagai batek beste osagaiaren kategoriako edozein hitzekin osatutako konbinazioen batez besteko portaera. Esaterako, *liburua irakurri* konbinazioaren portaera *liburua*+`aditza` konbinazio guztien portaerarekin konparatzea, eta orobat `izena`+*irakurri* konbinazioen portaerarekin.

- Bi portaeren arteko konparazioa egiteko erabiltzen den neurria. Aurreko bi portaeren definizioa hau da: hautatutako fenomeno morfosintaktikoen banaketa bana. Bi banaketen arteko distantzia neurtzeko, hainbat neurri garatu dira estatistikan.

Atal honetan, malgutasun morfosintaktikoa neurtzeko egin diren eragin handieneko ikerlanak aurkeztuko ditugu, hiru alderdi horietako bakoitzean erabili dituzten irizpide eta teknikak azalduz.

Barkema (1994a)

Ataza honetako lehen lanetako bat izan zen. Ingelesezko `noun+noun` konbinazioen hedapen morfosintaktikoak aztertu ditu. Lehenik, horrelako konbinazio-motek agertzen dituzten hedapen edo aldakuntza morfologiko eta sintaktikoen errepertorioa egiten du, konbinazioaren "oinarri-forma" edo konbinazio hutsa ere (*base form*) errepertorioko item bat dela. Esaterako, III.5 taulan *cold war* elkartearen hedapen batzuk ageri dira.

base form	*Cold War*
premodifying adjective	*renewed Cold War*
postmodifying prepositional phrase	*The Cold War between Nature Conservancy Council and the farmers*
premod. adjective + postmodifying clause	*The ackward cold war thought up by the American paranoids*

III.5 Taula: *cold war* elkartearen hiru hedapen (Barkema, 1994a: 43).

Ondoren, konbinazio aztergai bakoitzaren malgutasun-profila zehazten du, alegia, errepertorioko hedapen bakoitzaren maiztasun absolutuen (agerpenen) eta erlatiboen banaketa. Komeni da argitzea maiztasun erlatiboen batura 1 dela, hau da, errepertorioko hedapenek paradigma bat osatzen dutela, eta ezin direla batera gertatu testuko agerpenetan.

Azkenik, banaketa hori erreferentzia-banaketa batekin konparatzen du; erreferentzia osatzeko, `noun+noun` konbinazioen profilen batezbestekoa kalkulatzen du. Hedapen bakoitzerako behatutako maiztasun erlatiboaren eta erreferentziako balioaren arteko kenduraren bidez, batez besteko profilerako distantzia portzentuala adierazten du Barkemak. Hedapenik gabeko konbinazio hutsaren distantzia portzentuala zenbat eta handiagoa izan, hainbat

eta malgutasun txikiagoko konbinazioa dugu (hau da, konbinazioak erreferentzia-portaerak baino gutxiagotan onartzen ditu hedapen morfosintaktikoak). Esaterako, *cold war*en kasuan, oinarri-formaren maiztasun erlatiboa 0,8952 da, eta batezbestekoaren profilean, berriz, 0,3964. Diferentzia +0,4988 da, eta hortik ondorioztatzen du Barkemak *cold war* malgutasun txikikoa dela.

Wulff (2008)

Barkemaren bidea garatu duen egile nabarietako bat da. Ingelesezko `V-NP` konbinazioak aztergaitzat harturik (*draw X line, make X point, break X heart*...), haien aldakuntzen oso azterketa zabala eta zehatza egiten du, hiru eratako malgutasunak bereizirik[8]:

- sintaktikoa: perpaus bat patroi sintaktiko desberdinetan agertzea (pasiboa, erlatiboa, galde-perpausa...).
- lexiko-sintaktikoak: `V-NP` egituraren kanpo- eta barne-modifikazioak (hedapenak eta txertatzeak). Esaterako, izenaren modifikatzaileak (adjektiboak, izenak, preposizio-sintagmak...), adberbioak...
- morfologikoa : izenaren flexioa (determinatzaileak, plurala...) edo aditzarena (aspektua, aldia, modua, pertsona...).

Wulffek ohar garrantzitsu bat egiten du hiru malgutasun horien aldakuntza-parametroen izaera dela eta. Malgutasun sintaktikoan, perpaus bat aldakuntza-mota jakin batean sailka daiteke, hau da, aldakuntza-parametro bakarra dago, eta kasu bakoitzean balio bat hartzen du. Beraz, aldakuntza sintaktikoek paradigma itxia osatzen dute (Wulff, 2008: 91). Aldakuntza lexiko-sintaktikoak ez, ordea; edo ez behintzat Wulffek definitu dituen moduan (hau da, aldakuntza-parametro bakunak definitu ditu, perpaus jakin batean konbinaturik ager daitezkeenak: adjektiboa bakarrik erants daiteke, preposizio-perpausa edo izena, baina horien konbinazioak ere bai (izena eta adjektiboa, izena eta preposizio-perpausa, eta abar). Azkenik, malgutasun morfologikoan, bereizitako aldakuntza-parametro bakoitzaren balioek paradigma bat osatzen dute; esaterako, aditzaren aldiak (*tense*) *past/present/future/nonfinite* balioetako bat hartzen du perpaus bakoitzean; izen-sintagmaren numeroa (*number NP*), berriz, *singular/plural* balioak har ditzake.

[8] Wulffen terminologian: *tree-syntactic*, *lexico-syntactic* eta *morphological*.

Horiek horrela, Barkemaren moduko metodo bat aplika daiteke Wulffen malgutasun sintaktikoa (*tree-syntactic*) aldakuntza-parametro bakarraz neurtzeko, baina ez gainerako bi malgutasunen kuantifikazio globala egiteko. Horrelako neurri bat malgutasun lexiko-sintaktikorako erabiltzekotan, aldakuntza-parametro bakoitzari aplikatu behar zaio, bakoitzaren paradigma *aldakuntza/aldakuntzarik ez* banaketa delarik. Malgutasun morfologikoaren kasuan, aldakuntza-parametro bakoitzaren paradigmari aplika dakioke.

Bestetik, Wulffek Barkemaren metodoaren beste eragozpen batez ohartarazi gaitu: desbideratzearen noranzkoa (Wulff, 2008: 81). Izan ere, malgutasun-motaren arabera, ez da gauza bera aztergai dugun konbinazioa batez besteko portaera baino malguagoa edo zurrunagoa izatea. Axola zaiguna ez da zenbateraino urruntzen den, zein aldetara urruntzen den baizik. Bistan dena, bi horiek bereizi ezean, batezbestekoa baino malguagoak diren konbinazioak idiomatikotzat jotzeko arriskua dago.

Horrek guztiak eragin erabakigarria du malgutasuna kalkulatzeko metodoan eta erabil daitezkeen neurrietan.

Noranzkoak esanahirik ez duen kasuetarako, Wulffek bi proposamen egiten ditu:

- Lehena Barkemaren neurriaren hedapen bat da, eta horren helburua da batezbestekoarekiko (bere hitzetan, oinarri-lerroarekiko) desbideratze txikiek eragin txikia izatea malgutasunaren balio globalean, eta alderantziz. Horretarako, aldakuntza bakoitzaren behatutako maiztasun erlatiboaren eta itxarondakoaren arteko kendura erabili beharrean, Wulffek kendura horien karratuen batura (SSD, *sum of squared deviations*) eta horren bertsio normalizatua (NSSD) proposatzen ditu. Normalizazioa egiteko, emaitzetan ageri den SSD handienari 1 balioa ematen zaio, eta txikienari, 0.

- Bigarrena *relative entropy* izeneko magnitudea da. Informazioaren Teoriako entropiaren definizioa hau da:

$$H(p) = \sum_{x \in X} p(x) \log_2 p(x) \tag{III.47}$$

Eta honela definitzen du Wullfek entropia erlatiboa:

$$H_{rel} = \frac{H}{H_{max}}, \tag{III.48}$$

non H_{max} sistemaren entropia maximoa den, hau da, sistemaren egoera guztiak aleatorioki banatuta daudenean eta probabilitate berekoak direnean entropiak duen balioa:

$$H_{max} = \sum_{1,N} \frac{1}{N} \log_2 \frac{1}{N} = \log_2 N \tag{III.49}$$

Wulffek ez du adierazten, baina froga daiteke III.49 ekuazioa entropia erlatiboaren ohiko adierazpenaren baliokidea dela (Kullback-Leibler edo KL dibergentzia izenaz ere ezaguna dena):

$$H_{rel} = KL(P \parallel Q) = \sum_i p(i) \frac{\log_2 p(i)}{\log_2 q(i)} \tag{III.50}$$

Gure ustez, eztabaidagarria da erreferentzia-portaeratzat entropia maximoko banaketa har daitekeen, horrek ez baitu zertan izan aztergaitzat dugun konbinazio-motaren batezbesteko portaera, edo portaera prototipikoa. Esaterako, parametroa numeroa bada (*singular /plural*), entropia maximoko egoeran parametroaren balioak ekiprobableak dira, hau da 0,5. Baina litekeena da, esaterako, `izena+aditza` konbinazioetan, batez beste, izena maizago agertzea singularrean pluralean baino.

Azkenik, desbideratzearen noranzkoa esanguratsua den kasuetarako, *directional entropy* izeneko neurria sortu du Wulffek, guk *entropia direkzional* deituko duguna. Kalkulatzeko, lehenik, entropia erlatiboa unitatetik kentzen da $(1 - H_{rel})$, eta, bigarren urrats batean, noranzkotasuna esleitzen zaio: bigramaren aldakuntza-parametroa erreferentzia-portaera baino maizago gertatzen bada (hau da, malguagoa bada), zeinu positiboa esleitzen zaio, eta alderantziz. $[-1, +1]$ eskala bat osatzen da, malgutasun-mailaren araberakoa: zenbat eta balio negatibo handiagoa, hainbat eta malgutasun txikiagoa.

Wermter eta Hahn (2004)

Alemanezko `PNV` (`preposition-noun-verb`) konbinazioen aldagarritasuna aztertzen dute (*modifiability*). Corpus bateko `PNV` konbinazio guztiak erauzi ondoren, preposizioaren eta aditzaren artean gertatzen diren osagarri lexikal guztiak identifikatzen dituzte. Konbinazio bakoitzerako, osagarri bakoitza

ikusteko probabilitatea kalkulatzen dute, eta, konbinazioaren aldagarritasuna karakterizatzeko, probabilitate handieneko osagarriaren probabilitatea hartzen dute:

$$\mathcal{MOD}(PNV_{triple}) := \arg\max \mathcal{P}(PNV_{triple,Supp_k}), k = [1, n] \quad \text{(III.51)}$$

non:

$$\mathcal{P}(PNV_{triple,Supp_k}) := \frac{f(PNV_{triple,Supp_k})}{\sum_{i=1}^{n} f(PNV_{triple,Supp_i})} \quad \text{(III.52)}$$

`PNV` jakin bakoitzaren probabilitatea hau izanik:

$$\mathcal{P}(PNV_{triple}) := \frac{f(PNV_{triple})}{\sum_{j=1}^{t} f(PNV_{triple_j})} \quad \text{(III.53)}$$

Konbinazioaren kolokabilitatea (*collocability*) honela definitzen dute:

$$\mathcal{COLLOC}(PNV_{triple}) := \mathcal{MOD}(PNV_{triple}) \times \mathcal{P}(PNV_{triple}) \quad \text{(III.54)}$$

Fazly eta Stevenson (2007); Fazly et al. (2009)

`verb+object` konbinazioak idiomatikotasunaren lau propietateak neurtuz sailkatzeko lanean, finkotasun sintaktikoa neurtzean III.4 taulako patroiak kontsideratzen dituzte aldakuntzatzat.

Patroi horiek paradigma bat osatzen dute, eta egileek ohartarazten gaituzte boz pasiboaren kasuan determinatzaileak ez dituztela bereizi, pasiboarekin espero izatekoa bide den datu-barreiatzea saihestearren.

Konbinazio baten finkotasuna kuantifikatzeko, KL dibergentzia erabiltzen dute, era honetan:

$$\text{Fixedness}_{\text{syn}}(v, n) := D(P(pt|v, n)||P(pt)) = \sum_{pt_k \in \mathcal{P}} P(pt_k|v, n) \log \frac{P(pt_k|v, n)}{P(pt_k)} \quad \text{(III.55)}$$

Pattern No.	Pattern Signature			Example
1	v_{act}	det:NULL	n_{sg}	*give money*
2	v_{act}	det:*a/an*	n_{sg}	*give a book*
3	v_{act}	det:*the*	n_{sg}	*give the book*
4	v_{act}	det:DEM	n_{sg}	*give this book*
5	v_{act}	det:POSS	n_{sg}	*give my book*
6	v_{act}	det:NULL	n_{pl}	*give books*
7	v_{act}	det:*the*	n_{pl}	*give the books*
8	v_{act}	det:DEM	n_{pl}	*give those books*
9	v_{act}	det:POSS	n_{pl}	*give my books*
10	v_{act}	det:OTHER	$n_{sg,pl}$	*give many books*
11	v_{pass}	det:ANY	$n_{sg,pl}$	*a/the/this/my book/books was/were given*

III.4 Irudia: Finkotasun sintaktikoa neurtzeko patroiak (Fazly et al., 2009: 69).

$\mathcal{P}$ III.4 taulako patroi-multzoa da; $P(pt|v,n)$, konbinazioaren portaera, eta $P(pt)$, *verb-object* bikoteen portaera tipikoa (batez bestekoa). Beraz, Barkemak eta Wulffek NSSDrekin bezala, portaera orokorra erabiltzen dute erreferentziatzat.

$\text{Fixedness}_{\text{syn}}$-ez gain, Fazly eta Stevensonek (2007) beste neurri hauek ere erabiltzen dituzte:

$\text{Pattern}_{\text{dom}}$, konbinazio bakoitzaren patroi nagusia zehazteko:

$$\text{Pattern}_{\text{dom}} := \arg\max_{pt_k \in \mathcal{P}} f(v,n,pt_k) \tag{III.56}$$

$\text{Fixedness}_{\text{adj}}$, konbinazioaren izenak adjektiboa hartzeko duen finkotasun-maila neurtzeko:

$$\text{Fixedness}_{\text{adj}} := D(P(a_i|v,n)||P(a_i)) \tag{III.57}$$

Azkenik, konbinazioak *izenondoa bai/ez* aukeretatik nahiago duena zehazteko:

$$\text{Odds}_{\text{adj}}(v,n) = \frac{P(a_i = present|v,n)}{P(a_i = absent|v,n)} \tag{III.58}$$

Bannard (2007)

`verb+noun` egiturako konbinazioak aztertzen ditu, eta bera da erreferentzia-portaeratzat osagaien portaera erabiltzen duen egile bakarra. Argudioa da

konbinazioak baduela aldakuntzak izateko aurretiko probabilitatea, osagaien probabilitateetatik eratorria.

Zentzuzkoa dirudi Bannarden argudioak. `izena+aditza` osaerako konbinazioetan, izenak modifikatzaileak hartzeko izan dezakeen joera konbinazioko aditzak baldintzatua egon daiteke. Hori betetzera, litekeena da malgutasuna neurtzeko erreferentziatzat `izena+aditza` egituren batez besteko joera hartzea ez izatea emaitza adierazgarriak lortzeko prozedurarik doiena. Nolanahi ere, hipotesia da, eta esperimentalki egiaztatu behar litzateke, portaera orokorrarekin lortutako emaitzekin konparatuz, baina Bannardek ez du egin.

BNC corpusa RASP etiketatzailearen bidez prozesatu, eta aldakuntza hauek kontsideratzen ditu malgutasuna kuantifikatzeko:

- Determinatzaile bat txertatzea: *run the show* → *run their show*; *make waves* → *make more waves*
- Izen-sintagmaren barne-modifikazioa (esaterako, izenondoa sartzea izenaren eta aditzaren artean): *break the ice* → *break the diplomatic ice*
- Esaldia pasiboan agertzea: *call the shots* → *the shots were called by*

Malgutasuna neurtzeko, elkarrekiko informazio puntual baldintzazkoa izeneko neurria erabiltzen du (CPMI - *conditional pointwise mutual information*). z hitza emanik, x aldakuntza sintaktikoaren eta y hitzaren arteko PMIa ekuazio honen bidez kalkulatzen da:

$$I(x;y|z) = H(x|z) - H(x|y,z) = \log_2 \frac{p(x|y,z)}{p(x|z)} \tag{III.59}$$

Aldakuntzaren arabera, z eta y izena edo aditza dira. Esaterako, esaldi pasiboen kasuan, z aditza da, eta y izena. Izen-sintagmaren barne-modifikazioan, alderantziz. Malgutasunaren neurri global bat izateko, Bannardek aldakuntza sintaktikoen CPMIen batura egiten du.

III.9 Malgutasun lexikalaren neurketa

Malgutasun lexikala neurtzeko ohiko prozedura da konbinazioaren osagaietako bakoitzaren ordezkagarritasuna konputatzea, ordezkotzat osagaiaren sinonimoak, kuasisinonimoak edo semantikoki erlazionatutako hitzak erabiliz. Beraz, hori aurrera eramateko giltzarri da osagaien ordezkotzat erabil daitezkeen item lexikalen baliabideak eskuratzea. Bigarren, ordezkagarritasuna neurtzeko prozedura diseinatu egin behar da; hau da, ordezkoen bidez

osatutako "aldaeren" ezaugarriak jatorrizko konbinazioaren ezaugarriekin konparatzeko teknika bat behar da.

Lin (1999)

Arlo honetako lan bide-urratzailea izan da. III.7 atalean azaldu bezala, Linek konposizionaltasunik eza neurtzeko metodotzat aurkeztu zuen bere lana, baina, aipatu atalean azaldu genituen zenbait autoreren iritzia eta kritika kontuan izanik (Baldwin et al., 2003), atal honetan sailkatu eta aurkeztuko dugu.

Linek 125 milioi hitzeko kazetaritza-corpus batetik, honelako egiturako hirukoteak erauzten ditu: (`head type modifier`), non `type` osagaiak mendekotasuna adierazten duen. Esaterako, *John married Peter's sister* perpausetik, hirukote hauek ateratzen dira: (`marry V:subj:N John`), (`marry V:compl:N sister`) eta (`sister N:gen:N Peter`). Erauzi dituen 80 milioi mendekotasun-erlazioetatik esanguratsuak iragazi ondoren, thesaurus bat eratzen du, hirukote bakoitzerako honako antzekotasun-neurri hau kalkulatuz (Lin, 1998: 318):

$$\mathsf{PMI}(HTM) = \log \frac{P(A,B,C)}{P(B|A)P(C|A)P(A)} = \log \frac{|HTM| \times |*T*|}{|HT*| \times |*TM|} \qquad \text{(III.60)}$$

Kolokazio bat konposizionala den ebazteko, irizpide hau darabil Linek:

> «A collocation α is non-compositional if there does not exist another collocation β such that (a) β is obtained by substituting the head or the modifier in α with a similar word and (b) there is an overlap between the 95% confidence interval of the mutual information values of a and β.»

Pearce (2001)

Pearcek kolokazioak erauzteko proposatzen duen bidea azaldu berri dugun Linen ildo beretik dator, eta berak honela aurkezten du:

> «A pair of words is considered a collocation if one of the words significantly prefers a particular lexical realisation of the concept the other represents.»

Aukera lexikalen iturritzat, WordNet erabiltzen du. Kolokazio hautagai baten "indarra" (*collocation strength*) neurtzeko, hau proposatzen du:

$$s = \frac{f' - f''}{f'}, \tag{III.61}$$

non f' den konbinazio hautagaiaren w hitz batek synset bereko hitzekin dituen agerkidetza-maiztasunetatik handiena, eta f'', berriz, gainerako agerkidetzen maiztasun handiena.

Beraz, hitz batek "nahien" duen aukera gainerakoetatik zenbateraino nabarmentzen den neurtzen du. s neurriak $[0 - 1]$ arteko balioak hartzen ditu; zenbat eta handiagoa, hainbat eta indar handiagokoa kolokazioa.

Fazly eta Stevenson (2007); Fazly et al. (2009)

`verb+object` konbinazioen finkotasunaren bigarren osagaitzat ikertzen dute finkotasun lexikala, eta Linen (1999) ildo beretik ekiten diote. Hark sortutako thesaurus bera ere erabiltzen dute, baina finkotasuna neurtzeko beste bide bat proposatzen. Kontsideratzen dute konbinazio baten PMI neurriaren balioa zenbat eta gehiago aldendu ordezkatze lexikalaren bidez sortutako aldaeren PMIen batez besteko baliotik, konbinazioa lexikalki finkatuagoa dagoela, zurrunagoa dela. Aldentze hori neurtzeko, z neurria erabiltzen dute (III.62).

$$\text{Fixedness}_{\text{lex}}(v, n) := \frac{\text{PMI}(v, n) - \overline{\text{PMI}}}{s} \tag{III.62}$$

Van de Cruys eta Moirón (2007)

Nederlanderazko `verb+prepositional phrase` osaerako konbinazioekin egiten dute lan, eta, Linek bezala, ez-konposizionaltasun semantikoaren adierazgarritzat jotzen dute konbinazioaren osagai bat bere sinonimo edo kideko hitz batez ordezkatzerik ez izatea.

Alpino etiketatzailearekin prozesatutako 500 milioi hitzeko prentsa-corpus batetik, UF hautagaiak erauzten dira, eta gero matrize bat osatzen da, batetik maiztasun handieneko 5 000 konbinazioak eta bestetik maiztasun handieneko 10 000 izenak dituena.

Konbinazio baten izen baten ordezkoak eskuratzeko, antzekotasun distribuzionaleko teknikak darabiltzan *clustering*- edo multzokatze-prozesu bat inplementatu dute. Izenen bektoreen pisu edo balioetan, PMI neurria erabiltzen dute, maiztasun absolutuaren ordez.

Azkenik, konbinazio bateko izenak konbinazioko aditzerako[9] duen joera

[9] `verb` diotenenan, `verb+preposition` egituraz ari direla argitzen dute.

neurtzeko, KL dibergentzian oinarritutako neurri bat erabiltzen dute, hautespen-murrizketen arloko Resniken (1996) lanean inspiraturik. Lehenik, izenen banaketaren eta izenek aditz bat emanik duten banaketaren arteko KL dibergentzia kalkulatzen da (S_v):

$$S_v = \sum_n p(n|v) \log \frac{p(n|v)}{p(n)} \tag{III.63}$$

Ondoren, aditz batek izen jakin bakoitzerako duen joera ($A_{v \to n}$):

$$A_{v \to n} = \frac{p(n|v) \log \frac{p(n|v)}{p(n)}}{S_v} \tag{III.64}$$

Azkenik, $R_{v \to n}$ bi hauen arteko erlazio edo arrazoitzat definitzen da: aditz batek izen jakin batekiko duen joera, eta ordezkoen C multzoan dauden izenekiko dituen joeren batura. Hau da:

$$R_{v \to n} = \frac{A_{v \to n}}{\sum_{n' \in C} A_{v \to n}} \tag{III.65}$$

$R_{v \to n}$-ren balio handiak konbinazioaren finkotasunaren seinale dira. Konbinazioak ordezkorik ez duenean edo ordezkoak corpusean agerpenik ez duenean, ordezkoaren $p(n|v) = 0$, eta $A_{v \to n}$ zehaztugabea da. Kasu horretan, egileek 0 balioa esleitzen diote, eta, R indizea, beraz, 1 da.

Orobat $R_{n \to v}$, hau da, izen batek konbinazioko aditzerako duen joera, aditzen multzoko gainerako aditzekin konparatuta.

Egileek aitortzen dute emaitzak ordezkoen multzoen kalitatearekiko sentikorrak direla. Bestetik, ez dute adiera-desanbiguaziorik egiten. Haien emaitzak Fazly eta Stevensonen (2007) $\text{Fixedness}_{\text{lex}}$-en emaitzekin konparatu, eta hobetu egien dituztela adierazi dute.

III.10 Propietateen neurketen konbinazioa sailkatze-atazan: ikasketa automatikoa

Zenbait autorek aurreko lau ezaugarriak edo horietako batzuk neurtzeko teknikak batera erabili dituzte, neurri konbinatu haztatuak proposatuz edo ikasketa automatikoko teknikak aplikatuz.

Venkatapathy eta Joshik (2005) `VN` konbinazioen ezaugarri batzuk konbinatzen dituzte: kolokazio-atributuak eta testuinguru-atributuak. Lehenen artean, maiztasuna, PMI, antzeko kolokazioen PMI diferentzia txikiena, eta

aditzaren objektuaren maiztasun banatua daude; bigarrenen artean, kolokazioaren eta bere aditzaren arteko LSA antzekotasuna; eta kolokazioaren izenari dagokion aditzaren eta kolokazioaren arteko LSA antzekotasuna. SVM algoritmoa erabiltzen dute ikasketarako, eta balidazio gurutzatua ebaluaziorako. SVM-Light tresnaren bidez, ranking-funtzio bat sortzen da, eta horrela konparatzen dituzte sailkatzailearen rankinga eskuz landutako erreferentziarekin. Ranking-korrelaziorako, Pearsonen hein-korrelazioko koefizientea erabiltzen dute. Emaitza onenak atributu guztiak konbinatuz lortzen dira, baina ekarpen handiena kolokazioaren izenari dagokion aditzaren eta kolokazioaren arteko antzekotasunak egiten du.

Pecina eta Schlesingerrek (2006) elkartze-neurriak konbinatzen dituzte, sailkatzaileen bidez kolokazioak/ez-kolokazioak bereizteko[10]. Ideiaren motibazioa da AM batzuen balioak nahiko independenteak direla linealki konbinatuz emaitzak hobetu daitezkeela pentsarazteko. Sare neuronalak, Linear Logistic Regression, Logical Discrimination Analysis eta SVM algoritmoak erabiltzen dituzte. MAP (*mean average precision*) darabilte metrikatzat. Emaitza onenak neurona-sareekin (5 unitate) lortzen dira, eta Linear Logistic Regression algoritmoarekin ondoren.

Fazly eta Stevensonek (2007) idiomatikotasunaren lau propietateen jakintza erabiltzen dute, lau kategoria bereizteko (esapide idiomatikoak, aditz arineko eraikuntzak, izen abstraktuko kolokazioak, eta konbinazio literalak). BNCtik erauzitako ($f \geqslant 25$) eta prozedura batzuen bidez orekatutako 563 konbinazioren multzoa erabili dute; oinarrizko 12 aditz hauetako bat dute: *bring, find, get, give, hold, keep, lose, make, put, see, set, take*. Beste bi atributu hauek ere erabili dituzte: aditza, eta izenaren kategoria semantikoa (WordNeteko lehen adieraren hiperonimoa). Erabakitze-zuhaitzak erabiltzen dituzte (C5.0) ikasketa-metodotzat. Ezaugarri guztiak konbinatuz lortzen da emaitza onena (% 58,3ko zehaztasuna), eta ekarpen handiena malgutasun sintaktiko eta lexikalarekin lotutako ezaugarriek egiten dute (% 50,0ko zehaztasuna), AM eta antzekotasun distribuzionaleko ezaugarrien gainetik.

Lin et al.-ek (2008) *Logistic Linear Regression* eredua erabiltzen dute bost AMren balioak konbinatzeko (f, t neurria, LLR, MI eta χ^2). Kazetaritza-corpus batetik erauzitako hiru egituratako konbinazio-multzoetatik ausazko azpimultzo bana hartzen dute (5 000 hautagai): egiturak: `verb, OBJ, noun; noun, ATTR, Adj; verb, Mod, adv`. Ebaluazio-multzoak eskuz etiketatuak dira. Metrikatzat P eta R erabiliz, emaitzek erakutsi dute ezaugarri guztien konbinazioa dela onena, eta hurrena t neurria.

[10] Erabiltzen dituzten neurri guztiak ez dira, berez, AMak; esaterako, kosinu-antzekotasuna ere kontuan hartzen dute.

III.11 Laburpena

Kapitulu honetan, fraseologia konputazionalaren arloan UFak testutik automatikoki erauzteko eta karakterizatzeko egindako ikerketen berri eman dugu.

Lehenik, erauzte- eta karakterizazio-urratsak bereizi ditugu. Lehenaren helburua da UF izateko hautagaiak izan daitezkeen konbinazioak testuetatik eskuratzea. Horretarako alderdi erabakigarri nagusia da corpusaren prozesamendu linguistikoaren maila. Batetik, ezinbestekoa dirudi testuko hitz-formen gutxieneko informazioa izateak, hala nola lema eta flexioaren informazio morfologikoa, eta are argiagoa da hori hizkuntza flexibo eta eranskarien kasuan. Bestetik, sistema aurreratuagoetan, azaleko edo sakoneko analisi sintaktikoak emaitza hobeak lortzeko aukera eman dezakeela azaldu dugu. Bi eskakizun horiei erantzutea hizkuntza bakoitzeko HParen teknologien mende dago; gure ikergaian zenbaterainoko prozesamendua erabil dezakegun lan esperimentalaren diseinuan aztertuko dugu.

Karakterizazioaren xedea, berriz, hautagaiak bere idiomatikotasun-mailaren arabera automatikoki antolatzea da. Antolaera horren arabera, bi karakterizazio-ataza bereizi ditugu: ranking-ataza eta sailkatze-ataza. Batean zein bestean, emaitzak erreferentzia batekiko ebaluatu behar dira, ataza bakoitzaren araberako metriken bidez. Ranking-atazan, hein-korrelazioko koefizienteak erabiltzen dira, erreferentzian idiomatikotasun-continuuma bi kategoriatan baino gehiagotan banatuta dagoenean; erreferentzia bitarra bada, IR arloko doitasuna, estaldura eta F neurria dira ohikoenak.

Marko teorikoan definitutako idiomatikotasunaren lau propietateetako bakoitza neurtzeko garatu diren metodologiak deskribatu ditugu. Propietate bakoitza karakterizatzeko, hari dagokion edo harekin asoziatzen den fenomeno behagaia neurtzen da.

Instituzionalizazioaren edo idiosinkrasia estatistikoaren kasuan, UFaren osagaien agerkidetza-datuetatik abiatuz, konbinazioren elkartze-maila estimatzen da, zenbait eredu estatistikoren araberako elkartze-neurrien bidez.

Konposizionaltasun semantikoa karakterizatzeko, antzekotasun distribuzionaleko hainbat teknika aurkeztu ditugu, bigramaren eta haren osagaien testuinguruen arteko konparazioan oinarritzen direnak. Lehen faktore garrantzitsu bat da testuingurua nola definitzen den eta hura errepresentatzeko zer informazio erabiltzen den. Bost alderdi aztertu ditugu: helmena; testuinguru-hitzak agerkidetza hutsez edo erlazio sintaktikoaren arabera aukeratuko diren; kategoria guztietako hitzak ala eduki-hitzak bakarrik sartuko diren; eta errepresentazioan sartuko den hizkuntza-elementua zer informazioz hornitu edo nola karakterizatuko den. Bigarren alderdi gakoa da testuin-

gurua nola errepresentatzen den. Metodo erabiliena *hitz-espazioaren eredua* (WSM) da, IR arloan ezaguna den bektore-espazioaren ereduaren egokitzapena dena, eta bektoreen arteko zenbait antzekotasun-neurri erabiltzen dira. Bektoreen dimentsionaltasuna gutxitzeko SVD teknika (*balio singularretan deskonposatzea*) darabilen LSA sistema ere (*ezkutuko semantikaren analisia*) erabili da zenbait ikerkuntzatan. LSAren erabileraren inguruan ikuspegi desberdinak daudela ikusi dugu, eta kontuan hartzeko auzia da gure diseinu esperimentala egiterakoan.

Malgutasun morfosintaktikoa neurtzeko, konbinazioaren portaera morfosintaktikoak erreferentzia-portaera batekiko duen distantzia erabiltzen da. Portaera aldakuntzen banaketaren bidez adierazten da, eta konparazioa bi erreferentzia-motarekiko egin daiteke: egitura bereko konbinazioen batezbestekoa; edo konbinazioaren osagai batek beste osagaiaren kategoriako edozein hitzekin osatutako konbinazioen batez besteko portaera. Azkenik, ohartaraztekoa da propietate hau berezia dela, bi arrazoirengatik. Batetik, anizkoitza da, zeren portaera aldakuntza morfosintaktiko multzo baten bidez zehazten baita; planteatu behar da, beraz, nola osatuko den neurri bateratu edo konbinatu bat. Bestetik, kontuan hartzen diren aldakuntzen espektroa hizkuntza bakoitzaren ezaugarrien araberakoa da, eta berariazko lan bat egin behar da, hizkuntza bakoitzean malgutasunaren adierazle diren aldakuntzak aukeratzeko eta atzemateko.

Malgutasun lexikalaren neurketaren funtsa ordezkagarritasuna da. Hautagaiaren osagai bakoitza haren sinonimoez, kuasisinonimoez edo semantikoki erlazionatutako hitzez ordezkatu, eta osatutako konbinazio berrien ezaugarriak jatorrizko hautagaiaren ezaugarriekin konparatzen dira. Horretarako, ordezkoen errepertorioak behar dira, hiztegiak eta glosario distribuzionalak izan daitezkeenak, eta horien estaldura kritikoa izan daiteke esperimentuen emaitzetan.

Azkenik, propietateen banakako neurketen emaitzak ataza berean konbinatzeko saioak egin dira, neurri konbinatu haztatuak proposatuz edo, batez ere, sailkatze-atazan ikasketa automatikoko teknikak aplikatuz. Gure helburuetatik hurbilen dagoen ikerketa Fazly eta Stevensonena (2007) da, idiomatikotasunaren lau propietateen neurketen emaitzak erabili baitituzte sailkatze-atazan. Haien emaitzetan, malgutasun sintaktiko eta lexikalarekin lotutako ezaugarriek egiten dute ekarpen handiena, eta interesgarria da hori gure esperimentuetan ere gertatzen den aztertzea.

IV. KAPITULUA

Lan esperimentalaren diseinua

IV.1 Diseinu esperimentalaren elementuak

I.3 atalean ezarri ditugun ikertze-helburuak lortzeko eta, zehazki, 6. puntuan planteatu ditugun hiru ikergai espezifikoei erantzuteko, ebidentziak lortzeko bidea emango digun diseinu esperimental bat landu behar da, honako elementu hauez osatua:

1 Propietate bakoitzarekin lotuta dagoen behagai bana (gutxienez) aurkitzea, eta behagai horietako bakoitza neurtzeko estrategia zehaztea eta teknologia garatzea.

2 Behagai horien neurketen konbinazioa eginez edo neurketen emaitzak batera prozesatuz, karakterizazioa egiteko metodoa.

3 Esperimentuetako karakterizazio-atazak definitzea. III.4 atalean azaldu bezala, karakterizazioa bi atazatara bidera daiteke:

   3. 1 Ranking-atazan, UF hautagaiak idiomatikotasun-mailaren arabera ordenatuz, idiomatikotasunaren continuuma sortzea.

   3. 2 Sailkapen-atazan, UF hautagaiak espektro jarraituan bereizten ditugun kategorietan edo motetan banatzea.

4 Esperimentuetan lortuko ditugun emaitzak sistema ideal batek lortuko lituzkeen emaitzekin konparatu behar ditugu, eta, horretarako, ebaluazio-metodologia antolatu eta erreferentzia-baliabideak garatu behar dira.

Atal hau amaitzeko, gure ikergaia zedarritzea falta dugu. Izan ere, orain artekoan "UF hautagaiak", "UF-motak" aipatu ditugu, baina jakina da UFen egitura morfosintaktikoa askotarikoa dela. Bada, fraseologia konputazionalaren arloko erauzketa eta karakterizazioaren lehen urrats hauetan, gure lana euskaraz gutxi esperimentatu den konbinazio-mota batera mugatzea hobetsi dugu: `izena+aditza` konbinazio bitarrak.

Kapitulu honen hurrengo ataletan, planteatu berri ditugun asmoak zehaztuko ditugu, eta azkenekotik hasiko gara, unitate ikergaien zehaztapenetik, alegia.

IV.2 Unitate ikergaiak: `izena+aditza` osaerako konbinazioak

IV.2.1 Deskribapena

`izena+aditza` osaerako UFek oso multzo ugaria, bizia eta garrantzitsua osatzen dute. Lehen begiratu batean, iduri luke `izena+aditza` konbinazio bat zer den argitu beharrik ere ez dagoela, aski bailitzateke jakitea izen bat aditz batekin osatzen duen UFa dela. Dena den, literaturan ez da beti era berean ulertu, eta komeni da ikerkuntza honetan zein `izena+aditza` konbinazio ikertuko ditugun zehatzago azaltzea.

Zenbait ikerlanetan, `izena+aditza` osaerako konbinazioak definitzen direnean, sarritan izena subjektu- eta, batez ere, objektu-funtzioa duen osagaitzat hartzen da (Venkatapathy eta Joshi, 2005). Esaterako, *to pay attention, baisser la tête, romper el hielo*. Aldiz, izena preposizio-sintagma bat dutenei, hala nola *to take into account* modukoei, `verb+PP` (*prepositional phrase*) edo `verb+preposition+noun` izendapena eman ohi zaie (Van de Cruys eta Moirón, 2007).

Dena den, badira `izena+aditza` izendapenaren barnean egitura sintaktiko gehiago kontuan hartzen dituztenak. Alemanez, esaterako, Heidek (1998) dio `izena+aditza` kolokazioetan izena aditzaren subjektua, osagarria edo adjuntua izan daitekeela, eta Krennen (2008) iritziz, FVG *Funktionsverbgefüge* direlakoetan[1], izena preposizio-sintagma baten parte izaten da askotan. Gaztelaniaz ere, Koikek (1998), *colocaciones sustantivo-verbales* direlakoez diharduela, `verbo+preposición+sustantivo` egiturakoak ere sartzen ditu kategoria horretan. Aditz arinaren definizioa egitean (II.4 atala), adierazi dugu izena gehienetan aditzaren objektua izaten dela, baina zenbait autoreren arabera definizioa zabalagoa dela ere bai, eta, esaterako, preposizio-sintagmak ere ager daitezkeela.

[1] Ingelesezko *light verb constructions* eman ohi da termino baliokidetzat.

Horiek horrela, badirudi zentzuzkoa dela euskaraz ere ikuspegi zabala hartzea, eta ikertu nahi ditugun konbinazioetan izena aditzaren argumentu zein adjuntu izan daitekeela onartzea, are arrazoi handiagoz euskara hizkuntza eranskaria dela kontuan hartzen badugu. II.2.2.2 eta II.4 ataletan, hurrenez hurren, kolokazioen eta aditz-lokuzio (edo predikatu konplexuen) egitura sintaktikoak sailkatzeko eredu batzuk azaldu ditugu. Irizpide orokor gisa, esan dezakegu errepertorio horietan dauden egituretatik, oro har, aditz batekin izen-kategoriako lema bat konbinaturik ageri den konbinazio-motak sartuko liratekeela gure multzoan. Ondorioz, sailkapen horietako mota hauek ez ditugu kontuan hartuko:

- Zabalaren [XS+LOT/LAG] multzotik, sintagman AdjS balioa duten konbinazioak; adib., *posible izan, oker egon*[2].
- `adberbioa+aditza` konbinazioetatik, aditzondoa dutenak (hau da, adberbioa postposizio-sintagma ez denean); adib., *hegaz egin, korrika egin, igeri egin*. Multzo horretakoak dira eratorpen-atzizki baten bidezko XS duten Zabalaren [XS+Adi$_{arina}$] multzoko batzuk (*haginka egin*).
- `adberbioa+aditza` konbinazioetatik, postposizio-sintagmaren burutzat izen bat ez dutenak; adib., *aintzakotzat hartu*.

Beraz, *harira etorri, gogora ekarri, martxan jarri* moduko konbinazioak `izena+aditza` motakoak direla kontsideratuko dugu[3], postposizio-sintagma izen baten gainean eratua baita. Hala ere, adizlagun horietako batzuk lexikalizatutzat jo ohi dira, eta horren adierazgarria izaten da sarrera gisa agertu ohi direla hiztegietan; esaterako, *aurrera, gora, alferrik*... Horregatik, adizlagun lexikalizatuen tratamendua arazoa izan daiteke, zeren, etiketatze linguistiko automatikoan `ADB` analisia badago, aditz batekin osatzen dituzten konbinazioak ezin izango baititugu `izena+aditza` konbinaziotzat erauzi.

Postposizio-sintagmaren osaera dela eta, Zabalak (2004) dio [PS+Adi] motakoetan *-tzat*, *-n*, *-ra* eta *-tik* postposizioek parte har dezaketela. Urizarrek (2012), berriz, *aurrez ikusi* eta *begiz jo* adibideak sartu ditu. Gainera, gure atariko esperimentuetan, gutxienez kolokazioak diruditen honelako adibide hauek topatu ditugu: *muturreraino eraman, bururaino eraman, arduraz*

[2] Zabalak AdjS gisa tratatzen du *eske ibili* predikatuaren *eske* osagaia; dena den, itzulpen literaltzat 'petición andar' eman izanak iradoki digu beharbada IS analisia eman nahi zitzaiola; gainera, hiztegietan *eske*ri ez zaio izond. kategoria egokitzen; *Elhuyar Hiztegi*ak, esaterako, `adb.`, `post.`, eta `iz.` ematen ditu (`http://hiztegiak.elhuyar.org/eu/eske`)

[3] Zabalaren [SP+V] multzoa, Urizarren aditz-lokuzioen C multzoa (`adizlaguna/aditzondoa+aditza`), eta Urizarren kolokazio-egituren 3. multzoa.

jokatu, legez kanporatu, ilusioz bete. Beraz, a priori ez diogu murriztapenik ezarriko postposizio-sintagmaren kasu-markari.

Laburbilduz, IV.1 taulan eman dugu gure erauzketa- eta karakterizazio-atazetan jomuga izan ditugun `izena+aditza` osaerako konbinazio-moten errepertorioa.

1	`izena`$_{\text{subjektua}}$ + `aditza`
	1.1 Absolutibodun sintagma: *burua joan, eguzkia sartu, ardoa garraztu, esnea galdu*
	1.2 Ergatibodun sintagma: *gogoak eman, loak hartu, suak hartu, ilunak jo*
2	`izena`$_{\text{objektua}}$+`aditza`: *hanka sartu, zubiak eraiki, lan egin, min hartu, bizarra egin, itxurak egin, aukera eman, erabakia hartu, zarata atera, eskerrak eman, urratsak egin, adostasuna lortu, gola sartu, elkartasuna adierazi*
3	`izena`$_{\text{subjektuaren pred.}}$ + `aditza`: *beldur izan, falta izan, giro egon*
4	`izena`$_{\text{objektuaren pred.}}$ + `aditza`: *atsegin ukan/°edun, damu ukan/°edun*
5	`izena`$_{\text{datiboa}}$ + `aditza`: *edanari eman, bideari ekin, lanari lotu*
6	`izena`$_{\text{adjuntua}}$ + `aditza`: *mendean hartu, borrokan sartu, martxan jarri, adarretatik heldu, larrutik ordaindu, burutik kendu, harira etorri, gogora ekarri, aurrera eraman, zerbitzura egon, sareetara bidali, muturreraino eraman, aurrez ikusi, arduraz jokatu, oinarritzat hartu*

IV.1 Taula: Ikerketa honetako erauzketa- eta karakterizazio-atazetan jomuga izan ditugun `izena+aditza` osaerako konbinazio-moten errepertorioa.

Esan gabe doa, izenaren kasua `izena+aditza` konbinazioaren ezaugarri bat da, zeren `izena+aditza` bi lema jakinek unitate fraseologiko desberdinak osa ditzakete, izenaren kasua zein den. Horregatik, kasuaren informazioa ezinbestekoa da erauzten ditugun bigramen adierazpenean. `izena+aditza` konbinazioak lema-lema bikote hutsen bidez soilik adieraziko bagenitu, ez ginateke gai izango *kontuan hartu / kontu hartu*, *egunkaria irakurri /egun-*

karian irakurri eta beste hainbat bikote bereizteko.

IV.2.2 Forma kanonikoa

Ikusi berri dugu `izena+aditza` osaerako UFen zehaztapenean izenaren kasua nahitaez aintzat hartzekoa dela. Baina bada horretan zerikusia duen beste fenomeno bat. Zeren `izena+aditza` konbinazio bakoitzak halako malgutasun morfosintaktikoa du. UFak ere zenbait formaz ager daitezke testuetan, baina, eskuarki, "forma kanoniko" jakin batean formulatzen dira hiztegietan edo datu-base lexikaletan. Lan honetan, helburutzat hartu dugu gure erauzte-sistemak UF hautagaiak forma kanonikoan automatikoki formulatzea. Horretarako, aurrerago azalduko dugun bigrama-normalizazioaren prozedura erabili dugu (ikus V.3.2). Baina zer irizpideren arabera erabaki forma kanoniko hori automatikoki nola sortu? Hori da hemen landu nahi dugun gaia.

`izena+aditza` konbinazioen malgutasun morfosintaktikoaren alderdietako bat izenaren mugatasuna da, hau da, izenak zein mugatzaile daraman (mugagabean, mugatu singularrean, mugatu pluralean edo plural hurbilean dagoen)[4]. Bestetik, izena kasu absolutiboan duten UFetan, aurreko banaketa-eskemari partitiboa gehitu behar litzaioke, eta, kasu desberdina izanda ere, konbinazio kanoniko beraren aldakuntzatzat hartu[5]. Bestela esanda, *ez zuen erabakirik hartu* perpauseko *erabakirik hartu* bikotea *erabakia hartu* forma kanonikoarekin erlazionatu beharko genuke. Horrelako UFetan maiz gertatzen den fenomenoa da izenak berezko mugatzaile bakarra izatea, baina partitiboa ere onartzea. Esaterako, *hautsak harrotu* konbinazioa *hautsik harrotu* forman ere erabil daiteke. Horrelakoak dira *lan egin* edo *min hartu* moduko aditz-elkarte asko.

Mugatasuna ez da malgutasun-alderdi bakarra, jakina, baina bada UFaren forma kanonikoa zehaztean eragin handiena duena. Mugatzailearen malgutasunarekin zuzenean erlazionatuta dagoen beste alderdi nagusia determinatzailearena da, mugatzailea determinatzaile-mota bat baita. Erabat malguak diren konbinazioen kasuan, agerikoa da hori, baina, kolokazioen artean ere badira horrelakoak: *urratsak egin, urrats bat egin, urrats hori egin*...; *gola sartu, hiru gol sartu, gol asko sartu*... Nolanahi ere, kontuan hartuta kasu gehienetan flexioa dela forma kanonikoa erabakitzeko erabiltzen den informazioa, eta, ikerketaren urrats honetan, bigramen erauzketara mugatu

[4]Zehazki, bi parametro daude hor jokoan: mugatasuna (*mugatua/mugagabea*) eta, mugatuaren kasuan, numeroa (*singularra/plurala/plural hurbila*). Adierazpen laburragoa erabiltzearren, *mugatasun* terminoa erabiliko dugu bi parametro horiek globalki aipatzeko.

[5]Sareko Euskal Gramatika. Mugatzaileak. http://www.ehu.es/seg/morf/5/3/2

dugula aztergaia, izenaren mugatasunean oinarritu dugu forma kanonikoaren sorrera[6]. Aipatu determinatzailearekiko malgutasuna VII.1.3 azpiatalean landuko dugu, malgutasun morfosintaktikoa neurtzeko kontuan hartuko ditugun alderdietako bat izango baita.

Horiek horrela, izenaren mugatasunaren ikuspegitik egoera hauek bereiz ditzakegu:

- UF batzuk erabat finkatuta daude, mugatasunaren aldetik ez dira malguak, eta izenak beti mugatzaile bera du. Adibidez, *gogoak eman* (mug: -*a*), *eskuz aldatu* (mug: ∅), *kontu hartu*, (mug: ∅), *besoetan hartu* (mug: -*ak*). Esan gabe doa zein den honelakoen forma kanonikoa: testuetan ageri den bakar hori bera.

- Beste muturrean, izena "aske" da edo, bestela esanda, konbinazio libre batean bezalatsu ager daiteke mugatuta. Esaterako:

  (13) **Gola** sartzen ahalegindu da.
  Partida berean hiru **gol** sartzeari deritzo *hart-trick*a.
  Horrela jarraituz gero, ez du **golik** sartuko.
  Nork sartu ditu gaurko bi **golak**?

  (14) **Erabakia** hartu dut azkenean.
  Hari dagokio **erabakiak** hartzea.
  Ez zuen **erabakirik** hartu.
  Zenbait **erabaki** hartzeko elkartu gara gaur hemen.
  Hizkera estandar jasorako, beraz, honako **erabakiok** hartu ziren[7].

  Izena mugatasunaren aldetik librea edo murrizketarik gabea denean, forma kanonikoa mugatu singularrean eman ohi da. Horretara, aurreko kolokazioen forma kanonikoak *gola sartu* eta *erabakia hartu* dira, hurrenez hurren.

- Bi horien artean, era askotako malgutasun-mailak izan ditzakegu. Horietan ez dugu konbinazio libreetan espero dugun flexio-banaketa estandarrik edo batez bestekorik. Esaterako, *kontuan izan* UFaren aldaera bat *kontutan izan* da, baina **kontuetan izan* ez da erabiltzen. Orobat *auzira/auzitara jo* (baina ez **auzietara jo*). Kasu tipiko bat lehen aipatutako aditz-elkarteak (*lan egin*, *min hartu*) edo izena generikoki erabiltzen den konbinazioak dira (*atentzioa eman, bostekoa*

[6] Erabaki horrek *alde bat utzi* edo *alde batera utzi* moduko konbinazioak aztergaitik kanpo utziko ditu.

[7] *Euskara batua*, Koldo Zuazo, Elkar, 2005.

eman). Euskarri-aditz edo aditz ahulen konbinazioak dira denak ere. Horrelakoetan, plurala eta mugagabea nekez aurkituko ditugu. Bestetik, *hautsak harrotu* edo *eskerrak eman* moduko esapideetan izena absolutiboaren mugatu pluralean eta partitiboan erabil daiteke.

Hiztegietako forma kanonikoetan gehien ageri diren kasuak mugatu singularra, plurala eta mugagabea dira, eta testuetako aldakuntzen maiztasuna da, antza, alderdi erabakigarria. Partitiboa, banaketa-eskema honen kide bat dela kontsideratu dugun arren, ez da hiztegietako formetan erabiltzen[8].

Beraz, badirudi lehen hurbilketa batean forma kanonikoa maiztasun handieneko mugatasun-aldakuntzaren araberakoa izatea dela aukera egokiena, baina, aldakuntza hori mugatu singularra ez denean, kontuz aztertu behar litzateke horren nagusitasunaren neurria. Litekeena da mugatu plurala edo mugagabea izatea maizkoena, baina esanguratsuak ez diren arrazoiengatik. Esaterako, izena zenbatzaile zehaztugabe batekin agertzeko joera handia denean (hala nola, *lau gol sartu, zenbait gol sartu*), litekeena da *gol sartu* aldakuntza izatea ugariena, baina ez dirudi horrek forma kanonikoari eragin behar liokeenik, eta mugatu singularrean ematea dirudi zentzuzkoena. Forma kanonikoa automatikoki esleitzeko sistema eraginkor batek horrelako ñabardurak kudeatzeko gai izan behar luke, baina, horretarako, zailtasun batzuk gainditu egin behar dira (ikus bigrama-normalizazioaren prozedura V.3.2 atalean).

Azkenik, bada kudeatzen askoz ere zailagoa den auzi bat: izenaren mugatasunaren arabera, batzuetan UF beraren aldakuntzak ez baizik kategoria desberdineko konbinazioak ditugu. Esaterako:

(15) Ea, goazen –esan zuen kontua ekar ziezaioten **besoa altxatuz**[9]

(16) Irabazi izan ditu lasterketa batzuk, baina gurekin oraindik ez. Ea denboraldi honetan **besoak altxatzen** dituen[10]

Lehen adibidea konbinazio libretzat har liteke seguru aski, baina bigarrena, deskodetzen aski erraza den artean, 'lasterketa batean irabazi' adiera duen esapide idiomatiko figuratibotzat joko genuke. Baina adiera horretan

[8] Horrek ez du esan nahi ordea, batzuetan maiztasun handieneko aldakuntza ez denik. Aztergai dugun corpusean, adibidez, ikusi dugu *jaramonik egin* eta *ezustekorik izan* direla, hurrenez hurren, *jaramon egin* eta *ezustekoa izan* forma kanonikoen aldakuntza maizkoenak.

[9] *Gorde nazazu lurpean*, Ramon Saizarbitoria (Erein, 2000)

[10] http://paperekoa.berria.info/kirola/2010-02-04/025/004/euskaltel_euskadiko_txirrindulariak_gonzalez_de_galdeanoren_hitzetan.htm

(nahiz irabazleak batzuetan beso bakarra altxatu), izena pluralean dela lexikalizatu da (ez dugu uste "Ea denboraldi honetan besoa altxatzen duen" ohikoa denik).

Beste kasu bat *zubia eraiki / zubiak eraiki* eta antzeko bikoteena da. Banaketa erabat argia eta zehatza ez den arren, lehena konbinazio librearen forma kanonikoa dela esango genuke, mugatasun-aldakuntza estandarra onartzen duena. Bigarren forma kanonikoa, berriz, esapide figuratiboarena da, 'desberdinen arteko komunikazioa eta harremanak erraztu, bideratu' adiera duena, eta, nagusiki, pluralean erabiltzen dena[11].

Mugatasunaren araberako adiera-banaketa automatikoki hautematea ez dugu ikerkuntza honen helburutzat ezarri, eta etorkizuneko ikergaitzat dugu. Horrek berekin dakar, ezinbestean, horrelako fenomenoak ezin izango ditugula bereizi, eta bigrama-normalizazioa egin ondoren, denak forma kanoniko beraren agerpentzat joko direla. Are zailtasun handiagokoa da, gainera, mugatasun berekoen arteko anbiguotasuna ebaztea (*Errekara joan eta aurpegia garbitu zuen / Neurri horiek hartuta, gobernuak aurpegia garbitu nahi du*). Ataza hori ikerkuntza honetatik kanpo dago.

IV.3 Idiomatikotasunaren osagai edo propietateak neurtzeko esperimentatu ditugun estrategiak

Ikerkuntza-lan honetan III.5 atalean deskribatu ditugun lau estrategiak aplikatu ditugu euskarazko `izena+aditza` konbinazioen karakterizazioa egiteko. Horretara, hauek dira gure esperimentazio-ildoak:

- Konbinazio hautagaien osagaien agerkidetza neurtzea, elkartze-neurriak erabiliz, idiosinkrasia estatistikoa edo instituzionalizazioa kuantifikatzeko.
- Konbinazio hautagaien eta haien osagaien arteko antzekotasun distribuzionala neurtzea, konposizionaltasun semantikoa kuantifikatzeko.
- Konbinazio hautagaien portaera morfosintaktikoa erreferentzia-portaera batekin konparatzea eta bien arteko distantzia neurtzea, mal-

[11] Segur aski, gaztelaniazko *tender puentes* esapidearen eraginez. Bide batez, erreparatzekoa da gaztelaniazko konbinazio librerako ez dela ohikoena *tender* aditza erabiltzea, *construir* baizik; euskaraz ere, *zubia egin* da konbinazio libre ohikoena (*Ibai gainean oinezkoentzako zubia egiteko proiektua*), nahiz eta *zubia eraiki* ere ageri den testuetan (*Erreken gaineko zubiak eraikitzen hasi ziren*); adiera idiomatikorako, berriz, *eraiki* aditzarekiko esapidea nagusitu da azkenaldian, izena, esan bezala, batez ere pluralean dela (*Belaunaldi zaharren eta berrien artean zubiak eraiki behar direla dio*).

gutasun morfosintaktikoa kuantifikatzeko. Erreferentzia-portaeratzat literaturako bi aukerak esperimentatu ditugu:

  - portaera orokorra: `izena+aditza` konbinazioen batezbestekoa.
  - osagaien portaera: `izena`+[`konbinazioko aditza`] konbinazioen eta [`konbinazioko izena`]+`aditz` konbinazioen batezbestekoak.

- Konbinazio hautagaien osagaien ordezkagarritasuna neurtzea, ordezkotzat sinonimoak edo antzekotasun distribuzional handieneko hitzak erabiliz, malgutasun lexikala kuantifikatzeko.

Lau neurketa horiek gauzatu ahal izateko, horietako bakoitzaren inplementazioaz gain, beste bi egiteko hauek zehaztu behar ditugu:

- Neurketak zein `izena+aditza` konbinaziori aplikatuko zaizkien, hau da, aipatu ditugun "`izena+aditza` konbinazio hautagaiak" nola erauziko ditugun testuetatik.

- Neurketaren emaitza ebaluatzeko zer ataza antolatuko dugun, zer ebaluazio-metodologia erabiliko dugun eta zer ebaluazio-erreferentziarekiko konparatuko diren emaitzak.

Esperimentuak azaltzeari ekin aurretik, hurrengo kapitulu banatan jorratuko ditugu azken bi egiteko horiek.

V. KAPITULUA

UF hautagaiak erauztea

V.1 Corpus-baliabideak

UFen erauzketa eta karakterizaziorako egin ditugun esperimentu guztietan, estatistikoki estimatu da konbinazio hautagaiaren ezaugarri bakoitzari dagokion magnitudea, eta horrek berekin dakar tamaina handiko testu-bildumak erabili behar izatea, estatistikoki adierazgarriak diren laginak prozesatuko badira. Ondorioz, ahalegin handia egin dugu ahalik eta corpus handiena erabiltzeko.

Ikerkuntza honetan, bi testu-bildumaz osatutako kazetaritza-corpusa erabili dugu:

- *Euskaldunon Egunkaria*: 2001-2002 bitarteko zenbakiez osatua. Hitz-kopurua: 27 849 883.
- *Berria*: Internetetik eratu dugun 2006-2010 bitarteko zenbakiez osatutako corpusa. Hitz-kopurua: 47 188 290.

Beraz, corpusak 75 038 173 hitz ditu guztira, eta euskaraz prozesatu den corpus handienetako bat da.

Corpusaren tamaina handia eta prentsak ohiko lexikoan duen eragina kontuan hartuta, corpus egokia eta interesgarria da. Dena den, kazetaritza-arlokoa izateak ondorioak ditu erauzten diren `izena+aditza` konbinazioen ezaugarrietan eta tipologian. Izan ere, komunikabideetan, eta zehazki, prentsan erabiltzen den fraseologia eta, esaterako, hizkuntza orokorrean, literaturan edo zientzian erabiltzen dena, ez dira bat bera, konbinazio batzuk

testu-motarekiko idiosinkratikoak direlako, eta estatistikoki ere erabileradatu desberdinak agertzen dituztelako (Altzibar, 2005: 4). Horrenbestez, kazetaritza-corpus batetik UFen erauzketaren bidez lortzen diren errepertorioak ez dira, *stricto sensu*, hizkuntza orokorraren adierazgarri. Nolanahi ere, proiektu honen helburua teknologia garatzea da, ez baliabide lexikal jakinak eratzea. Landu ditugun teknikak eta garatu ditugun tresnak beste ezaugarri batzuk dituen corpus bati aplika dakizkioke, eta corpusen horren ezaugarrien araberako UF-errepertorioak lortuko dira.

V.2 Corpusaren prozesamendua

III.3 atalean UF hautagaien erauzketaz jardutean, adierazi dugu prozesamendu linguistiko minimoa lematizazioa eta etiketatze morfosintaktikoa dela, baina analizatzaile sintaktikoen informazioa izateak eragin positiboa izan dezakeela emaitzetan. Euskaraz badaude mendekotasun-analisia aurrera eramateko garatzen ari diren zenbait tresna; zehazki, Ixa taldea garatzen ari den Maltixa erabiltzeko aukera aztertu dugu (Bengoetxea eta Gojenola, 2010). Zenbait arrazoi direla medio, alde batera utzi behar izan dugu. Batetik, ingurune esperimentala prestatu zenean, behar dugun tamainako corpusa etiketatzeko abiadura-performantzia ez zen behar bezain handia; bestetik, garatze-fasean egonik, doitasuna ere ez zen behar den mailakoa.

Horiek horrela, Eustagger tresna erabili dugu corpusa lematizatzeko eta etiketatzeko, eta ez dugu analisi sintaktikorik egin.

V.2.1 Etiketatze linguistikoa: Eustagger

Corpusa linguistikoki prozesatu dugu, UPV/EHUko Ixa taldearen Eustagger etiketatzaile automatikoa erabiliz (Ezeiza et al., 1998; Alegria et al., 2002). Eustaggerren irteera "txertatua" erabili dugu. V.1 irudian, *Berriaper honetan ez nuen atzo horri buruzko aipamenik ikusi* perpausaren analisia ikus dezakegu. Eustagerrek EDBLko kategoria-sistema erabiltzen du. Sistema horren erreferentzia zehatza Urizarrek (2012) eman du bere tesiaren A. eranskinean (2-8 or.) [1].

Ez dugu erabili Eustaggeren hitz anitzeko unitate lexikalak (HAULak) etiketatzeko HABIL modulua (Alegria et al., 2004a). Tresna horren bidez, EDBLn errepresentatuta dauden HAULak detekta eta etiketa daitezke. Gure lan esperimentalaren helburua da, ordea, testuetatik UFak automatikoki

[1] http://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak/1342621075/publikoak/TESIA

```
/<Berripaper>/<HAS_MAI>/
   ("berripaper" IZE ARR @KM>)
/<honetan>/
   ("hau" DET ERKARR NUMS DEK MG DEK INE @ADLG)
/<ez>/
   ("ez" PRT EGI @PRT)
/<nuen>/
   ("*edun" ADL B1 NR_HU NK_NI ERL MEN ZHG
    @+JADNAG_MP_OBJ @+JADLAG_MP_OBJ)
   ("*edun" ADL B1 NR_HU NK_NI @+JADLAG)
/<atzo>/
   ("atzo" ADB ADOARR @ADLG)
/<horri>/
   ("hori" DET ERKARR NUMS DEK DAT NUMS MUGM @ZOBJ)
/<buruzko>/
   ("buruz" ADB ALGARR DEK GEL @IZLG> @<IZLG)
/<aipamenik>/
   ("aipa+men" ADI SIN ATZ IZE ARR DEK PAR MG @OBJ @SUBJ)
/<ikusi>/
   ("ikusi" ADI SIN AMM PART ASP BURU @-JADNAG)
/<.>/<PUNT_PUNT>/
```

V.1 Irudia: Eustagger etiketatzailearen irteeraren adibide bat.

erauztea, eta ikuspegi horretatik, UF hautagai guztien portaera behatzea interesatzen zaigu.

V.2.2 Eustaggerren irteeraren tratamendua

Corpus anotatu horretatik, bigramen erauzketarako beharrezkoa den informazioa aukeratzen da, Perl script multzo bat exekutatuz. IV.2 atalean azaldutakoa kontuan hartuz, izenen informazio hau interesatzen zaigu:

```
forma lema kategoria azpikategoria kasua mugatasuna
```

`IZE` kategoria duten analisietan, azpikategoriaren balioa beharrezkoa da izen bereziak (`IZB` edo `LIB` azpikategoria dutenak) gero bigrama-sorkuntzan kontuan ez hartzeko.

Aditzen kasuan berriz, printzipioz lema eta kategoria baino ez dugu behar, aditzaren flexioak ez baitu eraginik konbinazioaren forma kanoni-

koaren formulazioan. Gainerakoetan, kategoria baino ez dugu behar, hain zuzen ere, bigrama-sorkuntzan kategoria horretako tokenekin bigramarik ez sortzea kontrolatzeko; azpikategoria gorde dugu, baina gainerako informazioa ez (`0 0`).

Aurrerago azalduko dugunez, malgutasun morfosintaktikoaren neurketan (VII.1.3.3 atala) Eustaggerren irteera modu aberatsagoan, ez hain simplifikatuan, erabili beharra izan dugu.

Aurreko adibidetik informazio hau aukeratzen dugu:

```
berripaper berripaper IZE ARR 0 0
honetan hau DET ERKARR 0 0
ez ez PRT EGI 0 0
nuen *edun ADL B1 0 0
atzo atzo ADB ADOARR 0 0
horri hori DET ERKARR 0 0
buruzko buruz ADB ALGARR 0 0
aipamenik aipamen IZE ARR PAR MG
ikusi ikusi ADI SIN_PART BURU 0
```

Hala ere, `ikusi` hitzaren analisiak ikusarazten duenez, lehen urrats batean aditzaren zenbait informazio atxiki egiten ditugu analisi-katean. `ikusi` aditzaren kasuan, partizipio-forma duela gordetzen dugu (`PART`), eta aspektu burutugabearen forma duela (`BURU 0`). Horren arrazoia da partizipio batzuk, kategoriaren posizioan `ADI` balioa dutenak, ez direla berez adiztzat eraman behar bigrama-sorkuntzaren prozesura. Esaterako, *herrialde aurreratuak, gobernuaren aliatuak* erako konbinazioak, berez, `izena+izenondoa` eta `izenlaguna+izena` konbinazioak dira, hurrenez hurren, baina Eustaggerren irteeran, *aurreratu* eta *aliatu* lemek `ADI` kategoria dute, eta, horretan geratuko bagina, `izena+aditza` eran tratatuko genituzke, eta, ondorioz, erauzketan zarata sartu. Adibidez, hona hemen *herrialde aurreratuen zerga-sistema* sintagmaren analisia:

```
/<herrialde>/
   ("herrialde" IZE ARR DEK ABS MG @OBJ @SUBJ @PRED)
/<aurreratuenen>/
   ("aurreratu" ADI SIN AMM PART DEK GEN NUMP
    MUGM DEK GEN NUMP MUGM @IZLG> @<IZLG)
   ("aurreratu" ADI SIN AMM PART DEK GEN
    NUMP MUGM ELI DEK GEN NUMP MUGM @IZLG> @<IZLG)
/<zerga-sistema>/
```

```
   ("zerga-sistema" IZE ARR MAR IZE ARR DEK ABS NUMs MUGM @OBJ
@SUBJ @PRED>)
```

Hortik, honelako analisiak sortzen ditugu lehen urratsean:

```
herrialde herrialde IZE ARR ABS MG
aurreratuenen aurreratu ADI PART GEN NUMP
zerga_m_sistema zerga_m_sistema IZE ARR 0 0
```

Horrek aukera ematen digu horrelakoak azaleko morfosintaxi-arauen bidez berretiketatzeko, eta zarata murrizteko. Gakoa horien analisian ageri diren `ADI PART` etiketak dira, baita aurretiko izena kasu eta mugatasun analisirik gabea izatea (`0 0`), edo absolutibo mugagabea izatea (`ABS MG`). Berez, `herrialde` tokenaren `ABS MG` analisia ez dagokio jarri dugun adibideari, honelako bati baizik: *zenbait herrialde bisitatu ditu*. Dena den, adibideak berak salatzen duenez, etiketatzaileak ez ditu beti bi analisi-aukera horiek ondo desanbiguatzen, eta, estaldura handitze aldera, horrelako testuinguruetan ageri diren `IZE ARR ABS MG` analisiak ere tratatu ditugu.

Informazio hori prozesatuz, honelako irteera lor dezakegu, aditzari `ADJ ADI_PART` analisia emanez:

```
herrialde herrialde IZE ARR ABS MG
aurreratuenen aurreratu ADJ ADI_PART GEN NUMP
zerga_m_sistema zerga_m_sistema IZE ARR 0 0
```

Beraz, horrelakoetatik ez da `izena+aditza` bigramarik sortuko. Antzeko prozedura erabili dugu *gobernuaren aliatuak* moduko segidetan, bigarrenaren analisia `ADI` bada, `IZE ADI_PART` analisiaz ordezkatzeko; hartara, `izena+izena` bigrama sortuko da gero, eta ez `izena+aditza` bigrama.

Azkenik, aditzen partizipioaren gainean eraikitako izenlagunek (esaterako, *hartutako erabakien* kateak) tratamendu berezia behar dute. Argi dago *erabakia hartu* UFaren kasu bat dela, eta komenigarria dela UF horren agerpentzat kontatzea, estatistika adierazgarriak izango baditugu. Baina bistan da izenaren forma ezin dela *erabakien* izan, genitiboa (edo dagokion kasu-atzizkia) ez delako UFaren ezaugarri bat, ondorengo izenarekiko lotura baizik (esaterako, *hartutako erabakien ondorioak*). Horregatik, horrelako partizipioak `ADI` hutsaz markatu gabe, `ADIZL` etiketarekin markatu ditugu, aurrerago, normalizazioaren urratsean, modu berezian prozesatu ahal izateko.

V.3 izena+aditza osaerako konbinazio hautagaiak lortzea

V.3.1 Bigrama-sorkuntza

Testutik bigramak sortzeko, Ngram Statistics Package–NSP erabili dugu (Banerjee eta Pedersen, 2003; Pedersen et al., 2011)[2]. NSPk testua prozesatu ahal izateko, corpusa testu jarraituan antolatu behar da, lerrotan, eta zuriune arteko tokenak erabiltzen dira konbinazioak sortzeko. Horregatik, Perl script batez aurreko analisi-lerroak token bihurtzen ditugu, analisiaren elementuak azpimarraz lotuta. Esaterako, aurreko adibideko `aipamenik` tokena honela errepresentatzen dugu NSPrako corpusean:

```
aipamenik_aipamen_IZE_ARR_PAR_MG
```

Aditzaren kasuan, oraingoan bai, lema eta kategoria baizik ez dugu eramaten. Gainera, urrats honetan `ADI` eta `ADT` kategoriak ez ditugu bereizten; izan ere, maila honetan berdin zaigu *gogora ekarri zidan*, edo *gogora zekarkidan* den, bi agerraldi horiek unitate-forma kanoniko bakarrarekin erlazionatu behar genituzke. Hau da, aski dugu `ekarri_ADI` sortzea[3].

Gainerako kategorietako token guztiak `non_tok` gisa adierazten dira; horretara, NSPk ez du horiekin bigramarik sortzen, baina kontuan hartzen ditu bigramak sortzeko leihoaren tamainarako.

Azkenik, komeni da zenbait puntuazio-markaz bereizitako tokenen bigramak ez sortzea. Kontsideratu dugu esaldi-amaierako puntua, puntua eta koma, bi puntuak, eta galdera- zein harridura-marka bigramak sortzeko mugatzat jotzekoak direla; ez ordea koma eta Eustaggeren irteeran `BEREIZ` etiketa duten bestelako puntuazio-karaktereak. Beraz, NSPrako corpusean, Eustaggeren irteeran analisi hau duten tokenez bereizitako token-multzoak lerro banatan antolatu dira: `PUNT_PUNT, PUNT_PUNT_KOMA, PUNT_BI_PUNT, PUNT_GALD` eta `PUNT_ESKL`. Corpusa, beraz, horrelako puntuazio-markez bereizitako tokenez osatutako lerro-segida da.

Horiek horrela, atal honetako lehen adibidea honela doa bigrama-sorkuntzara:

[2] http://search.cpan.org/dist/Text-NSP

[3] Aurrerago, ordea, aditzaren informazio morfosintaktikoa kontuan hartu behar izan dugula ikusiko dugu, malgutasun morfosintaktikoaren parametro batzuk, erlatiboa esaterako, neurtu ahal izateko (ikus III.8 atala.)

```
berripaper_berripaper_IZE_ARR_0_0 non_tok non_tok non_tok non_tok
non_tok non_tok aipamenik_aipamen_IZE_ARR_PAR_MG ikusi_ADI
PUNT_PUNT
```

V.3.1.1 Erauzketa egiteko aldagaiak

NSPk erauzketa-parametro batzuk doitzeko aukera ematen du:

- `--window`: bi token bigramatzat (hau da, agerkidetzatzat) hartzeko distantzia. Distantzia eskuineranzkoa da beti, hau da, 1 denean, token batez eta haren ondorengoaz osatutako bigrama sortzen du; 2 denean, bi bigrama sortzen ditu, bata aurreko bera, eta bestea tokenaz eta haren eskuinerantz bigarrena denaz osatutakoa.
- `--remove`: maiztasun batetik beherako bigramak ez ditu sortzen.
- `--nontoken`: nontoken.regexp fitxategian dauden adierazpen erregularren araberako karaktere-segidak tokentzat ez hartzeko aukera.
- `--stop`: stop.txt fitxategian dauden adierazpen erregularren araberako tokenak bigrama-sorkuntzan ez sartzeko (esaterako, lehen aipatutako "non_tok" segidak).
- `--newLine`: esaldi bereko tokenen arteko bigramak bakarrik sortzeko (bestela esanda, bigramak sortzeko mugatzat lerro-hasiera eta -amaiera hartzeko).

Aurrerago deskribatuko ditugun esperimentuetan, leiho-zabaleraren eta maiztasun minimoaren balio batzuen araberako bigrama-erauzketak egin ditugu.

NSPren `count.pl` scripta erabil daiteke esaldikako corpusetik bigramak dokumentu batean sortzeko, baina corpus handiak prozesatzeko, egileek `huge.pl` erabiltzea gomendatzen dute, eta `--split` parametroak adierazten du corpusaren dokumentua zenbatetan zatitu behar den. Scriptak zati bakoitzean `count.pl` scripta exekutatzen du, eta ondoren zati guztien batura eta birkontaketa egiten du. Guk `huge.pl` erabili dugu, `--split` parametroaren 50 balioaz. NSPk formatu honetan sortzen ditu bigramak:

```
kontrola_kontrol_IZE_ARR_ABS_NUMS<>zorroztu_ADI<>30 1087 185
```

Azken hiru zenbakiak dira, hurrenez hurren, bigramaren maiztasuna, izen-unigramarena, eta aditz-unigramarena. Ohartaraztekoa da, urrats honetan, ez ditugula `IZE ADI` edo `ADI IZE` bigramak bakarrik sortzen, `IZE IZE` eta `ADI ADI` kategoria-konbinazioak ere sortzen dira. Hurrengo urrats batean, aztergai ez ditugun bigrama horiek iragazi egiten ditugu.

Lehen bigrama-erauzketa linealaren eta gordina egin ondoren, prozesu hauek inplementatu ditugu:

- **Konbinazio beraren ordena desberdineko bigramak konbinatzea**. Lehen erauzketan ordena desberdinean sortu diren baina berez bigrama berari dagozkion bigramak konbinatzea. `--window` parametroa azaltzean esan dugunaren arabera, NSPk lehen urratsean testuko ordenan sortzen ditu bigramak. Adibidez, horrelakoak:

  ```
  eskutik_esku_IZE_ARR_ABL_NUMS<>heldu_ADI<>274 1251 3639
  heldu_ADI<>eskutik_esku_IZE_ARR_ABL_NUMS<>58 2683 1013
  ```

  Beraz, beharrezkoa da UF beraren agerpenak diren ordena desberdineko bigramak konbinatzea eta maiztasun-kontuak batzea. Hori eginez, lortzen dugu hitz baten inguruko leihoaren zabalera ± 1 izatea. NSPko `combig.pl` scriptak egiten du batze-lan hori, baina maiztasun handieneko ordena aukeratzen du biak batzeko, eta horrek ez du bermatzen kategoria-ordena beti bera denik; horretarako, hurrengo urratsa da beharrezkoa.

- **Bigrama guztiak osagai-ordena berean jartzea eta `izena+aditza` osaerakoak ez direnak iragaztea**. Esan dugunez, `IZE ADI` zein `ADI IZE` ordenako bigramak izan ditzakegu, eta denak `IZE ADI` ordenan jartzen ditugu (`ADIZL IZE` ordenakoak ere alderantzikatu egiten ditugu). Bestetik, orain arteko erauzketa-emaitzetan, `IZE IZE` eta `ADI ADI` bigramak daude; horiek ez zaizkigu orain interesatzen, eta iragazi egiten ditugu.

- **Benetako `IZE ADI` ez diren konbinazio batzuk iragaztea**. Predikatu konplexu batzuen kasuan (adib., *nahi izan*), lehen osagaiak `IZE` analisia du (gogoan izan ez ditugula HAULak detektatu). Beraz, *ikusi nahi dut* modukoetan, *ikusi nahi* token-segidak `ADI IZE` kategoria-konbinazioa du. Baina, esan gabe doa, ez da benetako `IZE ADI` konbinazioa. Horien ondoriozko zarata kentzeko, beti `ADI IZE` ordenan

agertzen diren bigramak iragazi egin ditugu. Ordena ADIZL IZE bada (*hartutako erabakien ondorioak*), iragaztea ez da aplikatzen.

- **Osagaien maiztasunak birkontatzea**. Aurreko urratsetan egindako aldaketen ondorioz, bigramen osagaien maiztasunak birkalkulatu egin behar dira, erauzketa kontsistentea izan dadin.

V.3.2 Forma kanonikoa esleitzea: bigramen normalizazioa

Ikusi dugu bigramen formulazioan izenaren tokena (forma) sartu dugula (IV.2.1), eta erakutsi dugu UF batzuetan izenak forma desberdinak har ditzakeela, UFaren malgutasun morfosintaktikoaren arabera (IV.2.2).

Adibidez, V.1 taulan, *erabakia hartu* forma kanonikoaren aldakuntzak dira, bigrama-erauzketari begira mugatasunean soilik bereizten direnak, PAR kasuan dagoenean izan ezik:

erabakia har lezake	ABS NUMS
erabakiak hartzen ditu	ABS NUMP
zenbait erabaki hartu behar ditugu	ABS MG
erabakiok hartuko ditugu	ABS NUMP HURB
ez zuen erabakirik hartu	PAR

V.1 Taula: *erabakia hartu* forma kanonikoaren mugatasun-aldakuntzak.

Bada, UF hautagai bakoitzaren estatistika adierazgarriagoak lortzeko, eta hautagai bakoitzaren forma kanonikoa proposatu ahal izateko, aditz jakin batekin konbinaturik agertzen diren kasu bereko izen-formak normalizatu egiten ditugu, maiztasun handieneko forma eta mugatasuna esleituz, eta bigrama-kontuak birkalkulatuz.

Kasua absolutiboa denean (ABS), partitiboa ere (PAR) kontuan hartzen da normalizazioan, horien agerpenak ABSri esleituz, baina ez forma kanonikoa esleitzeko. IV.2.2 atalean azaldu dugunez, partitiboa, nahiz kasu batzuetan maiztasun handieneko aldakuntza izan, ez da erabiltzen forma kanonikoetan, eta, ondorioz, blokeatu egin dugu maiztasun handienekoa denean, eta bigarren maizkoena eman dugu forma kanonikoan.

Mugagabearen (*hiru gol sartu*) edo mugatzailerik ezaren kasuan (*atzerrian lan egiten du*), aldakuntza horiek forma kanonikoraino iristea maiztasun-datuen mende jartzea erabaki dugu. Mota batekoek zein bestekoek

dituzte bestelako aldakuntzak (*gol ederra sartu; lan gutxi egiten du*), eta arazoak daude *gola sartu* eta *lan egin* forma kanonikoetaraino enpirikoki iristeko[4].

Aldiz, *kontu hartu* eta *kontuan hartu* konbinazioak ez dira bakar batean normalizatzen, izenak ez baitaude kasu berean (lehena absolutiboan, eta bigarrena inesiboan).

Perl script batek detektatzen ditu zein bigrama diren elkarren artean normalizatzekoak. `izen_lema+izen_kasua+aditz_lema` gakoa erabiliz, bigrama bakarra sortzen du maiztasun handieneko bigramaren forma eta mugatasunarekin, eta, azkenik, bigramen maiztasunak batzen ditu, baita izen-unigramenak ere.

Adibidez, *egunkaria irakurri* eta *egunkarian irakurri* forma kanonikoetara honela iristen gara:

```
egunkarietan_egunkari_IZE_ARR_INE_NUMP<>irakurri_ADI<>31 74 1956
egunkarian_egunkari_IZE_ARR_INE_NUMS<>irakurri_ADI<>43 76 1956
egunkariak_egunkari_IZE_ARR_ABS_NUMP<>irakurri_ADI<>67 486 1956
egunkaria_egunkari_IZE_ARR_ABS_NUMS<>irakurri_ADI<>126 1003 1956
```

```
egunkarian_egunkari_IZE_ARR_INE_NUMS<>irakurri_ADI<>74 150 1956
egunkaria_egunkari_IZE_ARR_ABS_NUMS<>irakurri_ADI<>193 1489 1956
```

Horrez gain, ergatibo singularra → absolutibo plurala aldaketa ere egiten dugu `izena+aditza` lema-konbinazio bereko bigrametan, bi bigramen maiztasunen arrazoia 1:5 baino handiagoa denean. Izan ere, etiketatzaileak ez ditu beti ondo desanbiguatzen horrelakoak, eta maiz ergatibo singulartzat etiketatuak absolutibo pluralak izaten dira. Erabilitako arrazoia (1:5) aski kontserbadorea da, hau da, absolutibo pluralaren aldeko ebidentzia-proportzio handia eskatu dugu ergatibo singularraren analisia ordezkatzeko.

Beraz, hau da *erabakia hartu*ren normalizazioa:

[4]Aukera bat litzateke egiaztatzea, mugagabea maizkoena denean, kasu guztietan determinatzailedun sintagma batean gertatze den ala ez. Hala denean, litekeena da froma kanoniko egokiena mugatua izatea (*gol sartu*). Dena den, horretarako urrats honetan dugun baino informazio aberatsagoa behar da (azaleko sintaxiaren informazioa erabili behar genuke erauzketan), eta alde batera utzi dugu.

```
erabakia_erabaki_IZE_ARR_ABS_NUMS<>hartu_ADI<>2658 6329 88447
erabakiak_erabaki_IZE_ARR_ABS_NUMP<>hartu_ADI<>1632 2397 88447
erabakiak_erabaki_IZE_ARR_ERG_NUMP<>hartu_ADI<>88 141 88447
erabakirik_erabaki_IZE_ARR_PAR_MG<>hartu_ADI<>211 211 88447
```

```
erabakia_erabaki_IZE_ARR_ABS_NUMS<>hartu_ADI<>4589 9361 88447
```

IZE ADIZL konbinazioen kasuan (V.2.2 atalean aipatutako *hartutako erabakien ondorioak* kasua), erabaki dugu izenaren kasua absolutibora ekartzea, estatistikoki hori baita aukera ohikoena, nahiz eta jakin kasu batzuetan ez dela horrela (adibidez, *etorritako bidetik joan zen* esalditik *bidetik etorri* UFa atera behar genuke, eta ez **bidea etorri*).

VI. KAPITULUA

Ebaluazio-metodologia eta baliabideak

VI.1 Oinarrizko irizpideak

III.4 azpiatalean ikusi dugu zein diren ebaluazio-metodologia egiteko zehaztu behar ditugun alderdiak:

- Ebaluazio-lagina.
- Erreferentzia, edo *gold standard*a.
- Metrika, edo ebaluazio-neurrien sistema.

Alderdi horietako bakoitzerako hartzen den erabakiak zerikusia du diseinu esperimentalaren zenbait alderdirekin eta, batik bat, karakterizazio-atazaren izaerarekin, hau da, ranking- edo sailkapen-atazekin. Puntu garrantzitsu hauek hausnartu ditugu ebaluazio-sistema diseinatzerakoan:

- Gure helburua ez da hautagaien karakterizazio bitarra egitea (UF bai/ez), edo ez hori soilik behintzat. UF-kategoriak bereizi nahi ditugu.
- Euskaraz UF-kategoriak (esaterako, esapide idiomatikoak eta kolokazioak) bereizten dituen iturri lexikografikorik ez dagoenez, behartuta gaude adituen eskulanaren bidezko ebaluazio-erreferentzia bat eratzera.

- Egin beharreko eskulana arrazoizkoa izan dadin, ebaluazio-erreferentzia erauzketaren azpimultzo bat izatea komeni da, hau da, ebaluazio-lagin bat osatzea. Ahal den neurri handienean, azpimultzo hori aleatorioa izatea da egokiena.

- Sailkapen-ataza ikasketa automatikoaren bidez eraman ohi da aurrera, eta teknologia horrek sailkatu behar den item bakoitzari buruz zenbait atributu edo ezaugarri (*feature*) erabiltzen ditugunean lortzen ditu emaitza onenak. Beraz, idiomatikotasuna neurtzeko tekniken emaitzak konbinatzea da horren oinarria.

- Interesgarria da teknika horien banakako ebaluazioa eta haien arteko konparazioa egitea, zenbait arrazoirengatik; interes intrintsekoaz gain, praktikoa izan daiteke, esaterako, sailkapen automatikoko sistemara atribututzat eraman nahi ditugun neurrien aurrehautaketa bat egiteko. Banakako emaitzen ebaluazioa egiteko, ordea, sailkapen-ataza ez da egokiena. Horretarako, ranking-ataza erabil liteke.

- Ranking-atazan, badirudi idiomatikotasun-eskala bat den erreferentzia batekin konparatu beharko genituzkeela emaitzak, eskala hori continuumaren adierazgarria dela onartuz. Badira bide horretatik jo duten lanak (McCarthy et al., 2003; Biemann eta Giesbrecht, 2011), baina gehien-gehienetan, UF-zerrenda bat den erreferentzia erabili ohi da, eta, emaitzak ebaluatzeko, doitasunean (P) eta estalduran (R) oinarritutako neurri-sistema. UFen erauzketa hutsa, sailkapenik gabe, helburu duten esperimentuetan, prozedura hori aski da. Gure kasuan, ordea, bideren bat esploratu behar dugu teknika jakin bakoitzak UF hautagaiak kategoriaren arabera nola ordenatzen dituen ebaluatzeko, gero sailkapen automatikoaren atazan erabiliko dugun erreferentzia berarekin ebaluatu nahi badugu, hau da, itemak bi baino kategoria gehiagotan (UF bai/ez hutsaz haraindi) karakterizatuta dauzkan erreferentziarekin. Heinen korrelazio-koefiziente bat izan daiteke aukera bideragarriena.

- Idiomatikotasuna neurtzeko teknikaren arabera, laginean sartzen diren hautagaiek baldintza kuantitatibo batzuk bete behar dituzte:

  – Antzekotasun distribuzionala eta malgutasun morfosintaktikoaren neurketek hautagaien testuinguru-informazio aberatsa behar dute, hau da, hautagaien agerpen-kopurua txikiegia ez izatea eskatzen dute. Oso testuinguru gutxi izanez gero, nekez egin daiteke neurketa fidagarria (Sahlgren, 2006: 76-77). Horrek esan nahi du

maiztasun-atari bat jartzea komeni dela. Esaterako, Sahlgrenek, esperimentuaren arabera, 50 eta 20ko atariak erabiltzen ditu, McCarthy et al.-ek (2003), 20koa, eta Katz eta Giesbrechtek (2006), 30koa.

- Elkartze-neurriekin, atari hori apalagoa izan daiteke. III.6.2 azpiatalean azaldu dugunez, aski frogatuta dago $f < 3$ konbinazioen estatistikak ezin direla fidagarritzat hartu, eta $f \geqslant 5$ konbinazioak soilik daudela kuantizazio-efektuen eraginetik libre. Hortik gora, teorikoki behintzat, ez dago ataria handitzeko arrazoirik, baina hori egiteko arrazoi praktiko bat da elkartze-neurri batzuen portaera eskasa izaten dela f txikiekin (esaterako, MI), eta zarata handia sartzen dutela emaitzetan (Evert, 2005: 166-167). De facto, Evertek berak esperimentu-sail handi bat egiten du $f \geqslant 30$ atariarekin.
- Malgutasun lexikala neurtzeko, interesatzen zaigu ataria oso altua ez izatea; bestela, osagaiak ordezkatuz sortzen diren maiztasun gutxiko konbinazioen ebidentziarik ez dugu izango, eta ezin izango dugu ordezkagarritasunaren neurketa fidagarririk egin. Irtenbide posible bat da ordezkoen konbinazioen datuak ebaluazio-laginarena baino maiztasun-atari baxuagoko erauzketa batean bilatzea.

- Ikasketa automatikoko esperimentuetan hautagaiak sailkatzeko erabiltzen den laginean, sistemak ikasteko adinako kopurutan izan behar ditugu UF-kategorien instantziak. Aurreikuspena da erauzketan hautagai gehienak konbinazio libreak izango direla, eta kolokazioak ugariagoak izango direla esapide idiomatikoak baino. Beraz, moduren bat behar genuke ebaluazio-laginean UFak gutxiegi ez izateko. Ideia ez da lagin artifizial bat egitea, kategoria guztietatik kopuru bertsua duena, baizik eta, irizpide arrazoituetan oinarrituta, UFen dentsitatea nahikoa den lagin bat automatikoki moldatzea. Seguru aski, aurreko puntuan maiztasun-atari bat jartzeko egindako proposamenak lagun lezake maiztasun txikiko konbinazio libre ugari laginetik kanpo uzten, horrek estaldura gutxitu lezakeela onartuta. Hautagaien erauzketan lortzen ditugun konbinazioen multzoak sailkapenean bereiziko ditugun kategorien arabera duen banaketa aurrez jakiterik ez dagoenez, ideia bat izan daiteke hiztegi-erreferentzia bat erabiltzea multzo horren atariko karakterizazio bat egiteko, eta gure helburuetarako egokiena den lagina lortzeko parametroak zehazten hasteko.

Aurreko gogoetagai guztiak kontuan izanik, jarraian zehaztuko dugun ebaluazio-prozedura diseinatu dugu. Lau ataletan banatu dugu: hiztegi-erreferentzia, ebaluazio-lagina, ebaluaziorako erreferentzia eta ebaluazio-metrika.

VI.2 Hiztegi-erreferentzia

Iturri hauetan agertzen diren `izena+aditza` osaerako unitateak bildu dira:

- HB: Euskaltzaindiaren *Hiztegi Batua*[1].
- EH: Ibon Sarasolaren *Euskal Hiztegia* (Sarasola, 1996).
- ELH: Elhuyar Fundazioaren *Euskara-Castellano/Castellano Vasco Hiztegia*[2] (Elhuyar, 2006).
- Intza: *Intza proiektua*[3].
- EDBL: Ixa taldearen EDBL datu-base lexikala[4] (Aldezabal et al., 2001).

Hiztegi horietako hitz anitzeko sarrera/azpisarrera guztiak aztertu dira, eta `izena+aditza` osaerakoak banan-banan aukeratu. Emaitza 3 720 UFren bilduma da. VI.1 taulan, iturri bakoitzak hiztegi-erreferentzian dituen UFen kopurua dugu.

Iturria	**UF-kopurua**
HB	1 121
EH	580
ELH	2 305
Intza	1 422
EDBL	810

VI.1 Taula: Hiztegi-erreferentzian dauden iturri bakoitzeko UFen kopuruak.

[1] `http://www.euskaltzaindia.net/hiztegibatua`; `http://www.euskaltzaindia.net/eaeb` gunetik deskargatua (2010-06-04ko bertsioa).
[2] 2010-06-02ko bertsioa.
[3] `http://intza.armiarma.com` (2010-06-02ko bertsioa).
[4] `http://ixa2.si.ehu.es/edbl` (2010-06-22ko bertsioa).

VI.2 taulan, UF kopuruak eman ditugu, iturri-kopuruaren arabera. UF gehienak iturri bakarretik jaso ditugu, eta oso txikia da iturri guztietan argitaratu diren UFen proportzioa.

Iturriak	**UF-kopurua**
5	94
4	332
3	339
2	468
1	2 487
totala	**3 720**

VI.2 Taula: UF-kopuruak, iturri-kopuruaren arabera.

Bildumaren osaera dela eta, gaingiroki esan dezakegu bi unitate-mota atzeman ditugula batez ere: esapide idiomatikoak eta morfosintaktikoki idiosinkratikoak diren *lan egin, min hartu* eta kideko aditz-elkarteak. Aldiz, kolokazio sintaktikoki malguak eta esapide figuratibo berriak gutxi ageri dira, eta, beste hizkuntza batzuetan gertatzen dena euskaraz ere gertatzera, hipotesia da testuetan horrelakoak ugariagoak izango direla hiztegi-erreferentzian baino (proiektu honen beharrezkotasunaren argudio bat da hori, hain zuzen ere). Beraz, ebaluazioan erreferentzia hau erabiliz gero, egiazko negatiboak ugariak izango lirateke, eta emaitzen ebaluazioa alboratua legoke.

Nolanahi ere, erreferentzia hau erabili dugu aurretiko erauzketa-esperimentu batzuen ebaluazioa egiteko, ebaluazio-lagina zehazteko interesgarriak diren datu batzuk eskuratzeko bidea ematen baitigu. Hurrengo atalean azalduko dugu lan hori.

Hiztegi-erreferentzia honen beste alderdi interesgarri bat da eskuz landutako laginarekin konpara daitekeela, eta bere osaerari buruzko datuak atera ditzakegula. Horrek baieztatu edo ezeztatuko digu aurreko paragrafoan egindako aurreikuspena, eta, baiezkoan, tesi honetan xedetzat hartu dugun egitekoak hiztegigintzari egin liezaiokeen ekarpenaren lehen zantzu bat izan genezake.

VI.3 Ebaluazio-lagina

Ebaluazio-lagin egokienaren ezaugarriak zehazteko, V.3.1.1 azpiatalean aipatutako leiho-zabaleraren (w) eta maiztasunaren (f) parametroen zenbait balio-konbinazioren araberako erauzketak egin ditugu, eta hiztegi-erreferentziarekiko ebaluatu. VI.3 taulan bildu ditugu horien inguruko datuak (bigrama-normalizazioa egikaritu ondorengo emaitzak dira).

erauzketa	**bigramak**	**hiztegi-erref.**	**%**
$w = \pm 1,\ f \geqslant 3$	147 553	1 054	0,71
$w = \pm 1,\ f \geqslant 30$	13 661	512	3,75
$w = \pm 1,\ f \geqslant 50$	7 969	424	5,32
$w = \pm 5,\ f \geqslant 10$	197 805	1 440	0,73
$w = \pm 5,\ f \geqslant 30$	44 830	666	1,49

VI.3 Taula: Leiho-zabaleraren (w) eta maiztasunaren (f) zenbait balio-konbinazioren araberako erauzketetan lortutako bigrama-kopuruak, eta hiztegi-erreferentzian dauden bigramen kopuruak, bigrama-normalizazioaren ondoren.

Emaitza horiek lehen iruzkin hauek iradoki dizkigute:

- Maiztasun txikiko atariak estaldura handitzeko aukera ematen du, baina kostua handia da, hau da, doitasuna oso baxua da. Esaterako, lehen bi erauzketak konparatuta, UFen ia erdiak (512) f-ren araberako rankingaren lehen hamarrenean kontzentratzen dira.
- Leihoa zabala izateak zarata handia sortzen du. Konparatu, esaterako $f \geqslant 30$ ataria duten bi erauzketak; $w = \pm 5$ zabalera duen erauzketan, 30 000 mila hautagai inguru gehiago daude, baina hiztegi-erreferentziako 154 UF gehiago baino ez.

VI.1 atalean adierazi dugu, idiomatikotasunaren osagai batzuk neurtzeko, hautagaiaren testuinguru-informazioa prozesatu behar dela, eta horrek berekin dakarrela agerpen-kopuru minimo bat duten bigramez osatu behar izatea ebaluazio-lagina. Bada, horri gehitzen badiogu maiztasun gutxiko UFak oso barreiatuta ageri direla maiztasun gutxiko hautagaien rankingean, zentzuzkoa iritzi diogu gure ebaluaziorako $f \geqslant 30$ duten erauzketak erabiltzeari. Balio hori Sahlgrenek (2006) eta McCarthy et al.-ek (2003) antzekotasun distribuzionaleko esperimentuetan erabilitakoen artekoa da (lehenak,

20 eta 50; bigarrenek, 29), eta, gainera, bat dator Evertek (2005) agerkidetzaren elkartze-neurriak ebaluatzeko erabilitako atariarekin.

Maiztasun minimo horrekin ($f \geqslant 30$), bigrama-erauzketa hauek prestatu ditugu, leiho-tamainaren eta bigrama-normalizazioaren eragina zehatzago neurtzeko:

- $w = \pm 1$ leiho-tamaina, bigramak normalizatuta
- $w = \pm 1$ leiho-tamaina, bigramak normalizatu gabe
- $w = \pm 5$ leiho-tamaina, bigramak normalizatuta

Normalizatu gabeko erauzketan, forma kanoniko beraren aldakuntzak hautagai dira (esaterako, *erabakia hartu* eta *erabakiak hartu*); hirugarrenean, non leihoa zabalagoa baita, bigrama hautagai gehiago ditugu, eta aztertu behar da hori zarata-eragilea den ala ez.

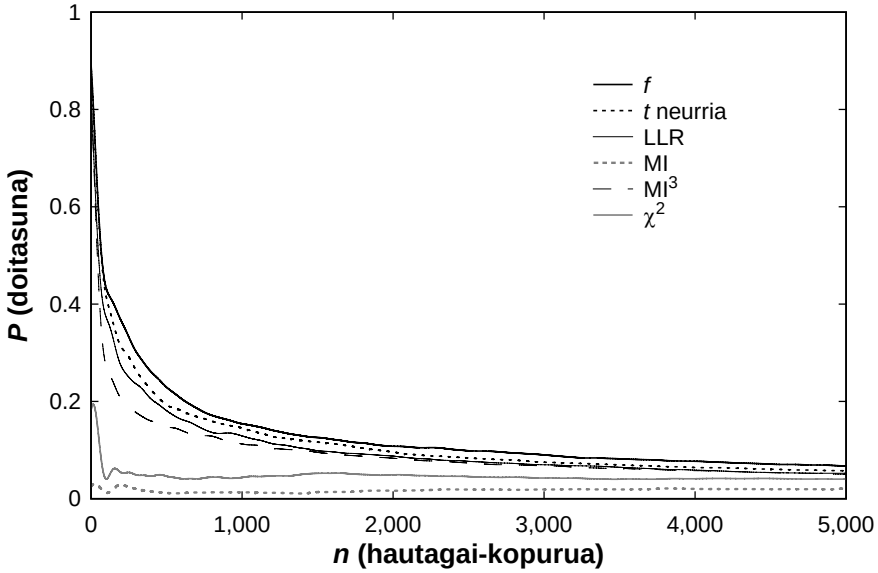
Bestetik, aipatu atalean ere azaldu dugu, ikasketa automatikoko esperimentuei begira, komenigarria dela sailkapenean erabiliko diren kategorietako instantziak oso desorekatuta ez egotea, hau da, kategoria batzuetako instantziak gutxiegi ez izatea. Baina badirudi hori gertatzeko arriskua dagoela baldin $f \geqslant 30$ baldintzapean egindako erauzketa edo horren ausazko lagin bat erabiltzen badugu ebaluazioan. Erauzketako 13 661 hautagaietatik 512 baino ez izatea hiztegi-erreferentzian aski proportzio txikia da, nahiz eta espero izatekoa den erauzketa eskuz aztertzean UF gehiago agertuko direla. Horrek pentsarazi digu moduren bat aurkitu behar genukeela UFen dentsitate handiagoko erauzketa-aldeez edo -zonez baliatzeko. Esan gabe doa, rankingaren hasierako partea da horretarako egokia. Baina f-ren araberako rankinga bakarrik erabiliko bagenu horretarako, argi dago neurri estatistiko baten arabera ariko ginatekeela ebaluazio-laginerako datuak hautatzen; ez litzateke miraria, orduan, elkartze-neurrien (AM) ebaluazioa egitean, neurri horren emaitzak izatea onenak.

Horregatik, erabaki dugu, erauzketa bakoitzaren ebaluazio-lagina osatzeko, f-ren bidezko rankingak ez ezik, beste zenbait AMren araberako rankingak ere erabiltzea. Beste AM hauek kalkulatu dira: t neurria (t-score), egiantz-arrazoiaren logaritmoa (*log-likelihood ratio* - LLR), MI, MI^3, eta khi karratua (χ^2). AMak kalkulatzeko, VII.1.1 azpiatalean zehatzago deskribatuko dugun Everten (2005) UCS toolkit Perl softwarea erabili dugu[5].

AM bakoitzak sortzen duen rankingetik lehen n hautagaien multzoak hartu, eta multzo horien bildura egin dugu. Zenbait proba egin ondoren,

[5] http://www.collocations.de/software.html

hobetsi dugu $n = 2\,000$ izatea. Kopuru hori, hein batean, arbitrarioa da. Dena den, VI.1 irudiak erakusten duenez, hautagai-kopuru horretara iristerako doitasun-balioak nahiko egonkortuta daude, eta zentzuzkoa dirudi horretara bitarteko hautagai-multzoak kontuan hartzeak.

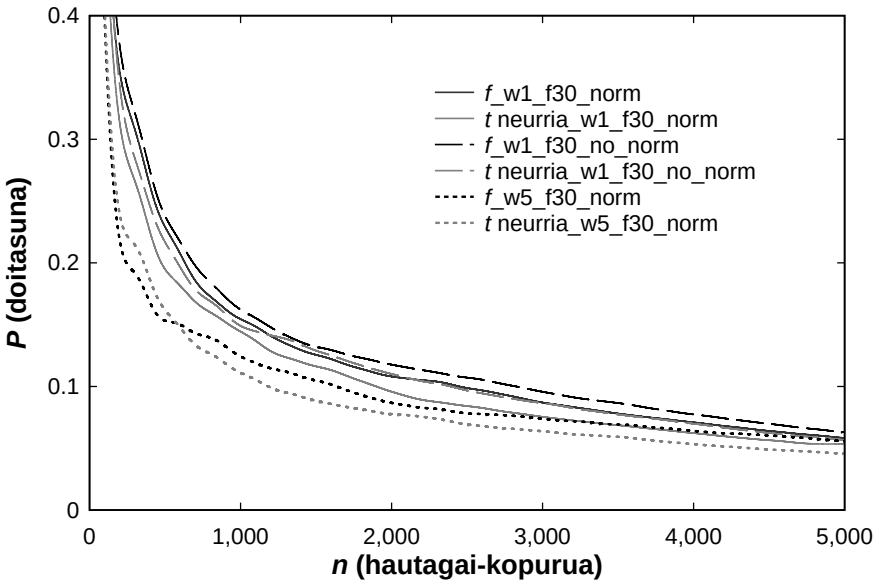


VI.1 Irudia: $w = \pm 1$ eta $f \geqslant 30$ parametroekin egindako erauzketaren lehen 5 000 hautagaien doitasun-emaitzak.

Lehen erauzketaren kasuan, bildura horrek 4 334 bigrama desberdin ditu; bigarrenak, 4 478; eta hirugarrenak, 5 386. Hiztegi-erreferentzian, hurrenez hurren, 250, 278 eta 263 UF daude. AMen balioen bidez sortutako rankingen aurreebaluazio bat egiteko, hiztegi-erreferentziarekiko doitasuna (P) eta estaldura (R) kalkulatu ditugu, rankingeko n hautagai onenen arabera.

VI.2 irudian, lagin horien hiztegi-erreferentziarekiko doitasun-kurbak ditugu. Emaitza onenak dituzten AMak baino ez ditugu bistaratu: f eta t neurria.

Emaitza onenak, nahiz t neurriarekiko aldeak esanguratsuak izan ez, f-ren bidezko rankingek dituzte, leiho txikia erabiliz eta normalizaziorik gabe. Leihoaren zabalera dela eta, eta hiztegi-erreferentziarekiko betiere, leiho zabala erabiltzea zarata-iturri da, doitasunari dagokionez behintzat.

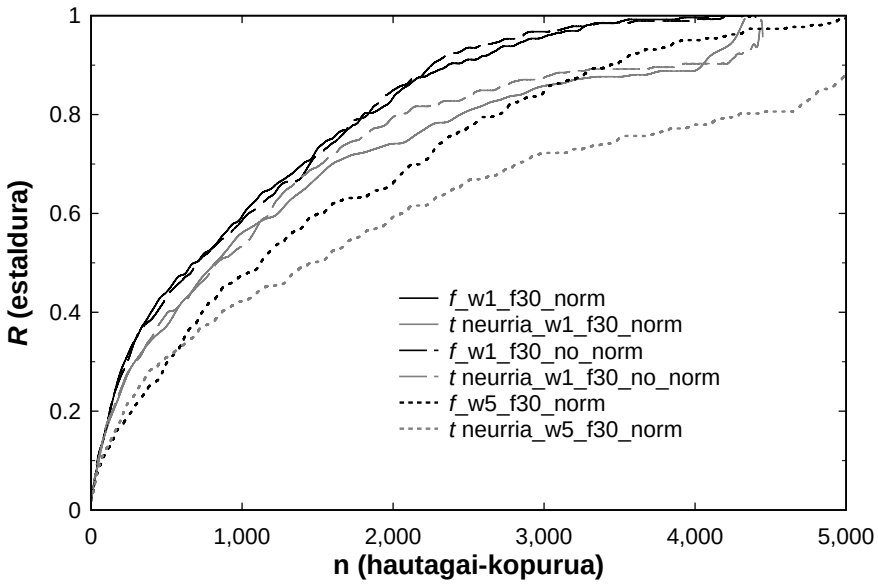


VI.2 Irudia: f-ren eta t neurriaren doitasun-kurbak hiru erauzketa-sorta hauetarako ($f \geqslant 30$): a) $w = \pm 1$, bigramak normalizatuta; b) $w = \pm 1$, bigrama-normalizaziorik gabe; and c) $w = \pm 5$, bigramak normalizatuta.

VI.3 irudian, berriz, f-ren eta t neurriaren araberako estaldurak ageri dira. Kasu honetan, normalizazioak, f-ren kasuan batez ere, ez du eragin nabarmenik.

Euskara ordena aski libreko hizkuntza izaki, espero izatekoa da, erabat finkoak ez diren UFen kasuan, leiho zabalaz egindako erauzketan bigramen agerpen gehiago lortzea; baina, aldi berean, erlazio sintaktikorik ez duten, eta, beraz, `aditza+izena` konbinazio baten osagaiak ez diren aditz eta izenen baterako agerpenak ere jasotzen ditugu, eta, bistan dena, hori zarata-sortzailea da. Badirudi, orduan, leihoa zabaltzeak, hautagaien erauzketaren fasean, ez duela ondoz ondoko osagaien erauzketak baino emaitza hobea ematen, ez behintzat osagaien mendekotasun sintaktikoaren informazioa ematen digun sistema bat erabili gabe.

Bestalde, bigrama-normalizazioak ere ez dio laguntzen doitasunari, nahiz eta normalizatu gabeko erauzketarekiko aldea txikiagoa den leiho zabalaren kasuan baino. Azter dezagun normalizatu gabeko erauzketan dauden hiztegi-



VI.3 Irudia: f-ren eta t neurriaren estaldura-kurbak hiru erauzketa-sorta hauetarako ($f \geqslant 30$): a) $w = \pm 1$, bigramak normalizatuta; b) $w = \pm 1$, bigrama-normalizaziorik gabe; and c) $w = \pm 5$, bigramak normalizatuta.

erreferentziako unitateetatik zein falta diren normalizatutako erauzketan. VI.4 taulan bildu ditugu unitate horiek. Haien eskuinean, normalizatutako erauzketan dagokion unitatea ageri da. Guztira, 24 UF dira.

Hiru azalpen-mota aurkitu ditugu (VI.4 taulan, eskuineko zutabean gehitu dugu azalpen-motaren zenbakia):

1 Hiztegi-erreferentziako forma kanonikoa ez da maiztasun handienekoa.

2 Normalizazio-prozesuaren ondorioz, bigrama ez da agertzen ezein AM-ren araberako rankingeko lehen 2 000 hautagaien artean (beraz, normalizatzean hautagaia ez da desagertu, baina rankingetan beherago agertzen da).

3 Eustaggerrek forma bati esleitutako lema ez da egokia, eta forma hori agertzen den bigrama berez ez dagokion normalizazio-prozesu batean

w1_f30_no_norm	**w1_f30_norm**	**Azalpen-mota**
ahaleginak egin	ahalegina egin	1
bide eman	bidea eman	1
bideak ireki	bidea ireki	1
dei egin	deia egin	1
eskuak garbitu		2
gaina hartu	gain hartu	1
galdera egin	galderak egin	1
galdera egin	galderak egin	1
hauts bihurtu		2
hitz eman	hitza eman	1
hitza izan	hitzak izan	1
iruzur egin	iruzurra egin	1
jabe egin		2
kalte eragin	kalteak eragin	1
kasu egin	kasurik egin	1
kontutan hartu	kontuan hartu	1
lan izan	lana izan	1
lanak izan	lana izan	1
lur jo	lurra jo	1
mehatxu egin	mehatxua egin	1
nahiago izan	nahi izan	3
nahiago ukan	nahi ukan	3
neurria hartu	neurriak hartu	1
partida ukan		2

VI.4 Taula: Bigrama-normalizazioaren ondorio batzuk ($w = \pm 1$ eta $f \geqslant 30$ parametroekin egindako erauzketa). Lehen zutabean, bigramak normalizatu gabeko erauzketan dauden eta normalizatutako erauzketan falta diren hiztegi-erreferentziako unitateak; bigarren zutabean, aurrekoei normalizatutako erauzketan dagokien forma kanonikoa; hirugarrenean, bat ez etortzearen azalpen-mota.

sarrarazten du. Esaterako, *nahiago ukan* bigramako *nahiago* tokenari *nahi* lema esleitu dio Eustaggerek.

Kasu gehienak lehen azalpenaren ondorio dira. *Kasurik egin* bigrama

normalizatuaren kasuan, badirudi irtenbide zentzuzkoena dela partitiboa ez izatea bigramaren forma kanonikorako aukeretako bat, eta ez sartzea horretarako maiztasunen konparazioan (agerpen-kontaketatik baztertu gabe). Besteetan, nabaria da izena mugagabean eman dela hiztegietan, eta behatu dugun testuetako erabilerak zer pentsatua eman lezake bigrama batzuen forma kanoniko egokiena hori ote den. Gainera, batzuetan beste fenomeno bat konbinatzen da: normalizatu diren bigrama batzuk UF desberdinak dira. Esaterako: *gaina hartu* eta *gain hartu* (berez, -*en gain hartu* dena).

Bigarren azalpenaren kasuan, eragile nagusia da normalizatu gabeko bigramaren formari dagokion lema maiztasun handikoa izatea, eta ondorioz, AMen balioak apalagoak izatea. Esaterako, *eskuak garbitu* konbinazio normalizatu gabean, `eskuak_esku_IZE_ARR_ABS_NUMP` unigramaren maiztasuna 275 da. Normalizatzean, ordea, `esku_IZE_ABS` gakoarekin bat datozen unigramen maiztasunen batura esleitzen zaio. Kasu honetan, balio hori 11 195 da; maiztasun marjinala handiagoa izanda, AMen balioa, oro har, jaitsi egiten da (bigrama gertatzea ez da lehen bezain "harrigarria", edo, bestela esanda, gertuago dago ausazko gertakari bat izatetik).

Azkenik, hirugarren kasuaren eragina gutxituz joango da etiketatzailean hobekuntzak egin ahala.

Aurreko azalpenak normalizatu gabeko erauzketarekin lan egiteko argudiotzat har litezke. Hala eta guztiz ere, kontuan hartu behar dugu:

- Egin dugun konparazioa kontrako noranzkoan eginez, normalizatu gabeko erauzketan ez dauden hiztegi-erreferentziako 22 UF agertzen dira normalizatutako erauzketan[6]. Kasu guztietan, normalizazioak, espero genuen bezala, forma kanoniko bereko aldakuntzen agerpen-maiztasunak batu eta bigrama igoarazten du rankingean. Har dezagun *denbora eman* kolokazioa adibidetzat. Normalizatu aurretik, bigrama hauek ditugu:

```
denbora_denbora_IZE_ARR_ABS_NUMS<>eman_ADI<>129 2255 115050
denborarik_denbora_IZE_ARR_PAR_MG<>eman_ADI<>65 363 115050
denbora_denbora_IZE_ARR_ABS_MG<>eman_ADI<>60 779 115050
```

  Normalizatu ondoren, berriz:

[6] Hauek dira: *argia ikusi, bainua hartu, bakea egin, defentsa egin, denbora egin, denbora eman, dirua xahutu, enkantean jarri, eskuz aldatu, falta egin, familia izan, isiltasuna hautsi, itxura hartu, itxurak egin, kiebra jo, kontuz ibili, larrua jo, lo egin, martxan ipini, opa ukan, premia ukan, puntan ibili.*

```
denbora_denbora_IZE_ARR_ABS_NUMS<>eman_ADI<>254 3608 110195
```

Lehen hiru bigrametan, forma kanonikoaren estatistikak "barreiatuta" daude, eta ez f-ren ez ezein AMren balioak dira nahikoak rankingeko lehen 2 000 hautagaien artean agerrarazteko. Normalizatutakoan, berriz, 1 782. postuan dago, f-ren araberako rankingean.

- Ebaluazio-lagina sailkatzeari begira, egokiagoa dirudi forma kanoniko beraren aldakuntzak bereiz ez agertzea. Izan ere, hizkuntzalarientzat, esaterako, *erabakia hartu* eta *erabakiak hartu* bi aldakuntzak sailkatu beharra bitxia izan daiteke, eta erabaki beraren bi gauzatze direla pentsa lezakete.

Horiek horrela, gure erabakia izan da leiho estuko ($w = \pm 1$) eta normalizatutako erauzketa (w1_f30_norm) erabiltzea eskuz sailkatuko den ebaluazio-erreferentzia lantzeko.

VI.4 Ebaluaziorako erreferentzia

Urrats honen helburua da idiomatikotasuna karakterizatzeko esperimentuen ebaluazioan erabiliko dugun erreferentzia sailkatua eratzea, aukeratu dugun erauzketako bigramak erabiliz. Aurreko atalean argudiatzen saiatu garenez, erauzketa horren ezaugarriak eta aurreikusten diogun UFen banaketa egokiak dira gure ikerkuntzaren helburuetarako. 4 334 bigramako bilduma da, eta eskura izan ditugun baliabideek ez digute aukerarik eman hautagai-kopuru hori osorik eskuz sailkatuko lukeen aditu-taldea antolatzeko. Hori dela eta, horren ausazko azpimultzo edo lagin bat aukeratu dugu automatikoki. Ausazko azpimultzo hori 1 200 itemekoa izatea erabaki da.

Hiru hizkuntzalariz osatutako aditu-talde batek egin du sailkatze-lana, eta ataza horren helburuak eta oinarrizko irizpideen inguruko informazioa eta argibideak jaso dituzte.

Sailkatu beharreko bigrama-laginean ikusi dugu hautagai batzuk ez direla benetako `izena+aditza` osaerako bigramak. Bi dira arrazoiak:

- Eustaggerren analisian esleitutako kategoria, `IZE` zein `ADI`, ez da zuzena. Esaterako, *egoki iritzi, gerora jakin* eta *larri zauritu* bigrametako lehen osagaiek `IZE` kategoria dute Eustaggerren analisian, eta, hurrenez hurren, `ADJ`, `ADB` eta `ADJ` kategoriak legozkieke. Halaber, multzo

honetakoak dira V.2.2 azpiatalean azaldu genituen *herrialde aurreratu* eta *gobernuaren aliatu* modukoak, non bigarren osagaiak, aditza gabe, izenondoa eta izena baitira. Horrelakoen bigarren osagaiaren `ADI` analisi desegokia zuzentzeko egindako tratamenduak ez ditu kasu guztiak konpondu, eta batzuk agertu dira sailkatzeko laginean (*jarrera kontrajarri* eta *politika bateratu*). `izena+izena` elkarte kasu bat ere badago (*bonba eraso*).

Sailkatze-lana egitean, horrelakoak `ez_NV` kategorian sailkatzeko eskatu zaie adituei.

- Bigrama beste unitate handiago baten parte da. Bi mota nagusi daude:
  - Erauzitako bigrama UF luzeago baten parte da. Esaterako, *hortzera erabili* dugu laginean, baina *hitzetik hortzera erabili* litzateke konbinazio osoa. Laginean ez dauden beste adibide batzuk ere aurkitu ditugu erauzketaren emaitzan; esaterako, *antzarrak ferratu* (→ *antzarrak ferratzera bidali*) eta *behera geratu* (→ *bertan behera geratu*).
  - Aurretik multzo ireki bateko elementu bat hartzen duten bi osagai baino gehiagoko unitateak: *aldiz irabazi* → determinatzailea + *aldiz bildu* (esaterako, *lau aldiz bildu, hainbat aldiz bildu...*); *andana bildu* → *jende/lagun/gazte...* + *andana bildu...*

  Horrelakoetarako, `ez_osoa` kategoria proposatu da.

Erabaki dugu horrelakoak ez kontuan izatea sailkatze-atazan eta etiketatzaileen arteko adostasunaren edo ITAren (*inter-tagger agreement*) kalkuluan. Horretarako, sailkatzaileen lehen lana izan da horiek detektatzea. Bakoitzak bi kategoria horietan sailkatuko lituzkeenak aurkeztu eta, eztabaidatu ondoren, adostasunez erabaki da 55 bigrama baztertzea eskuz sailkatu beharreko laginetik. Beraz, 1 145 bigrama sailkatu dira.

II.2.3 atalean sintagma-unitateetarako proposatutako sailkapena hartu dugu abiapuntuko jomugatzat. UFak bi kategoria nagusitan banatuta daude: esapide idiomatikoak (lokuzioak) eta erdiidiomatikoak, erabiliagoa den *kolokazio* terminoaz izendatu ohi direnak. Bina azpikategoria bereizi ditugu horietan. Beraz, hautagai bakoitza bost kategoria hauetako batean sailkatzea proposatu zaie adituei:

- esapide idiomatiko opakoa (`id_op`)
- esapide idiomatiko figuratiboa (`id_fig`)

- kolokazio murriztua (`col_res`)
- kolokazio irekia (`col_open`)
- konbinazio librea (`free`)

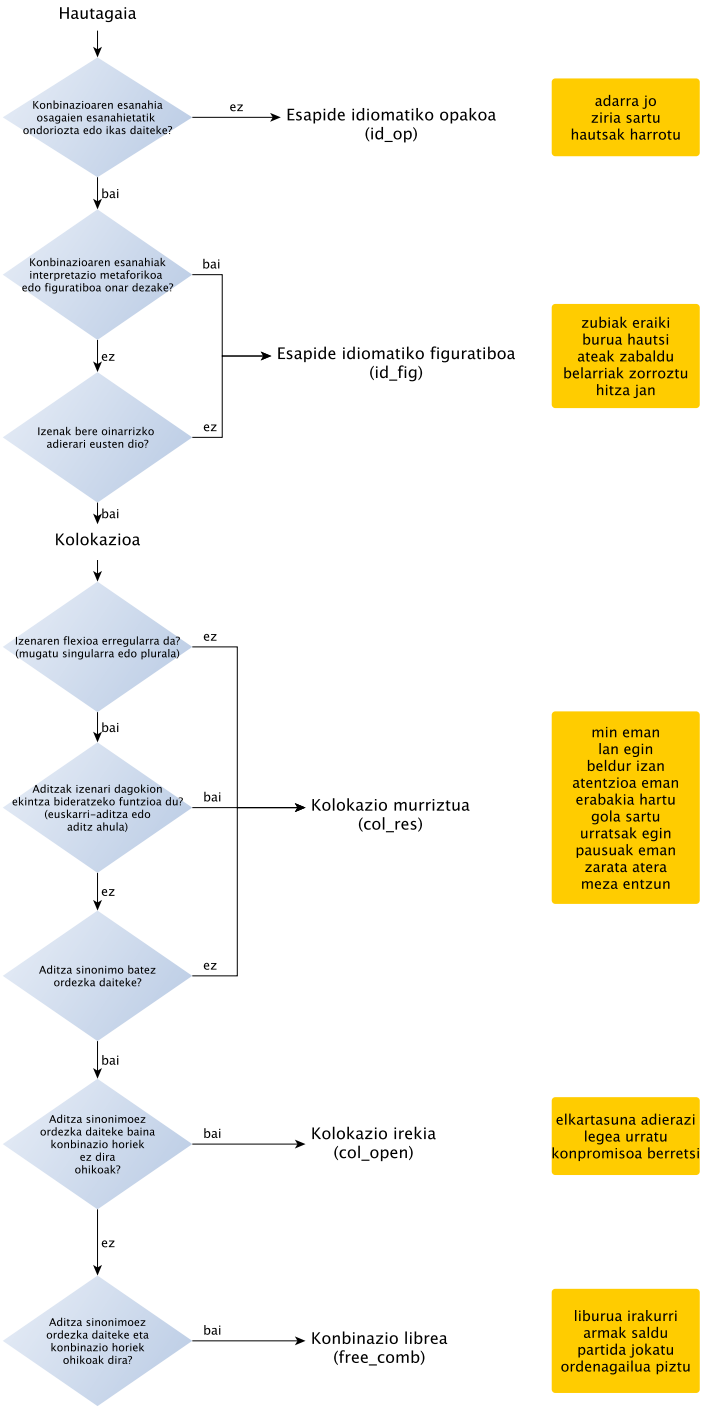
Sailkatze-prozesuan laguntzeko asmoz, VI.4 irudiko erabakitze-diagrama aurkeztu zaie adituei. Erabakitze-prozesua antolatzeko irizpideak Krennen (2004) lanean inspiratuta diseinatu ditugu. Lantaldea trebatzeko, lehen 100 hautagaiak sailkatu ditu aditu bakoitzak, eta emaitzak aztertu eta eztabaidatu ondoren, gainerakoak etiketatu dituzte.

ITA delakoa Fleiss kappa (κ) neurriaren bidez kalkulatu dugu (Fleiss, 1971). Adituen binakako adostasunak neurtzeko, berriz, Cohen κ erabili dugu. 5 kategoriako sailkapenerako ez ezik, 3 eta 2 kategoriako sailkapenen κ balioak ere kalkulatu ditugu. 3ko sailkapenean, esapide idiomatikoak sail bakarrera bildu ditugu, eta beste hainbeste kolokazioekin (`id`, `col`, `free` bereizketa). Azkenik, 2ko sailkapenean konbinazio libre ez diren kategoria guztiak sail berean bateratu dira (`MWE`, `free` bereizketa).

Hauek dira lortu ditugun emaitzak:

- 5 kategoria (`id_op, id_fig, col_res, col_open, free`)
  - Fleiss κ: 0,48
  - Cohen κ: 0,51 / 0,47 / 0,48
- 3 kategoria (`id, col, free`)
  - Fleiss κ: 0,55
  - Cohen κ: 0,62 / 0,51 / 0,52
- 2 kategoria (`MWE, free`)
  - Fleiss κ: 0,58
  - Cohen κ: 0,66 / 0,56 / 0,53

Espero izatekoa zenez, adostasuna apalduz doa erabilitako kategoria-kopurua handitu ahala. Landis eta Kochek (1977) emandako irizpideen arabera, Fleiss κ-erako lortu ditugun balioak adostasun ertainaren mailan leudeke (*moderate agreement*: 0,41 - 0,60), adostasun handiaren *substantial agreement* mailaren azpitik. Autore batzuek (Gwet, 2012) kritikatu izan dute horrelako irizpideak ematea, argudiatuz ez daudela arrazoi objektiboetan oinarrituta. Aitortu beharra dago UFen erauzketaren arloko zenbait lanetan adostasun-neurrien balio handiagoak argitaratu direla (Krenn



VI.4 Irudia: UFak sailkatzeko erabakitze-diagrama.

et al., 2004; Fazly eta Stevenson, 2007), baina baita gurearen antzeko emaitzak badirela ere, eta batzuetan, apalagoak. Esaterako, Kim eta Baldwinek (2007) % 51,71ko adostasuna lortu dute bi adituren artean ingelesezko 256 `VPC` (`verb-particle-construction`) lau kategoriatan sailkatzean. Bestetik, Pecinak (2010), hiru hizkuntzalarik egindako 6 kategoriako sailkatze-lan batean, 0,49ko κ lortzen du, eta, 2 kategoria soilik erabiliz, balioa 0,56ra igotzen da. Haren hitzetan, "This demonstrates that the notion of collocation is very subjective, domain-specific, and also somewhat vague." Vinczek (2012) 0,59ko kappa lortu du hiru adituk egindako ingelesezko unitateen sailkapenean, eta 0,65ekoa hungarierazkoan. Venkatapathy eta Joshik (2005), bi anotatzaile erabiliz, 0,61eko Cohen κ eman dute (3 kategoriarekin lortu dugun Cohen κ-ren balio baten mailakoa). Van de Cruys eta Moirónek (2007), hiru anotatzailek 200 konbinazio sailkatzean, Cohen κ-ren 0,60ko batez besteko balioa erdietsi dute. Azkenik, Street et al.-ek (2010), *American National Corpus*ean esapide idiomatikoak anotatzeko ataza batean, ITA balio baxu samarra lortu dute, Krippendorffen Alpha erabiliz: "The level of agreement among annotators regarding which phrases were actually idioms was no better than chance." Horrek guztiak neurri batean adierazten du ataza honen zailtasuna, eta UF izatearen edo idiomatikotasunaren kontzeptuaren irizpideak kasu errealei aplikatzeak sortzen dituen emaitzen barreiadura.

Emaitzak ikusita, badirudi, *substantial agreement* delakora gerturatzeko, 3 kategoriako sailkapen-eredua erabiltzea dela egokiena, kontuan izanik 2 kategoriakoak ez duela UF-kategorien arteko bereizketarik egiten. Gainera, badira horretarako beste arrazoi batzuk, sailkapen-lanaren emaitzek eta balorazioak erakutsi dizkigutenak:

- Esapide idiomatiko opakoen kategoria (`id_op`) oso instantzia gutxiri esleitu diete adituek (6, 15, 8). Esana dago lagina ausaz eratua dela, eta, alde batera, logikoa eta espero izatekoa da opakoen saila oso urria izatea, hala gertatzen baita hizkuntza-erabileran ere (Moon, 1998b: 79). Baina, hain adibide gutxi izanda, ikasketa automatikoko sistema batek nekez lor litzake taxuzko emaitzak. Horiek horrela, esapide idiomatikoen kategoria bakarra erabiltzea da zentzuzkoena.

- Kolokazio irekien kategoria oso labaina da, eta zalantza ugari sortzen ditu sailkatzaileen artean. Izan ere, osagaiak ordezkatzea onartzen duen konbinazio bat kolokazio irekitzat (eta ez konbinazio libretzat) jotzeko irizpidea idiosinkrasia estatistikoa da, eta anotatzaileak horretaz duen pertzepzioa da ebazpide bakarra (kontua ez baita ebaluatzaileak

corpuseko maiztasun-informazioa kontsultatzea), eta anotatzaileak ez dira oso seguru sentitu. Gainera, II.2.2 atalean azaldu genuenez, kolokazioen barneko bereizketa hori zalantzan jarri dute aditu batzuek. Horregatik, badirudi zuhurragoa dela sailkapena kolokazio-kategoria bakarrera mugatzea.

Beraz, hasieran genuen 5 kategoriako eredua alde batera utzi, eta 3 kategoriakoa onartu dugu ebaluazio-erreferentziarako: esapide idiomatikoak (`id`), kolokazioak (`col`) eta konbinazio libreak (`free`).

Esperimentuen ebaluazioan erabiltzeko, erabaki behar da adituen arteko erabateko adostasuna ez dagoen kasuetan item horiek erreferentzian mantenduko diren, eta, baiezkoan, zein kategoria esleituko zaion item bakoitzari. Hiru sailtzaileen adostasuna 776 bigramatan gertatu da; 365 bigramatan, bi sailkatzaile ados daude; azkenik, 4 bigrama kategoria desberdinean sailkatu ditu sailkatzaile bakoitzak. Azken sailkapenari begira, gure erabakia izan da item guztiak sartzea (1 145), 2 sailkatzaileren arteko adostasuna dagoenean kategoria automatikoki onartzea, eta gainerako kasuak (4) eztabaidatu eta adostasunez erabakitzea.

Azkenik, honela geratu da 1 145 itemeko ebaluazio-laginaren sailkapena: `id` 80; `col` 268; `free` 797.

VI.5 taulan, ebaluazio-ereferentziaren eta hiztegi-erreferentziaren arteko konparazioaren emaitzak daude. Hiztegian ez dauden eta sailkatzaileek UFtzat jo dituzten bigramen kopuruak baieztatu egiten du nolabait VI.2 atalean hiztegi-erreferentziaren estalduraz eta adierazgarritasunaz adierazi dugun irudipena. Eskuzko erreferentziako 348 UFetatik, 64 baino ez daude hiztegi-erreferentzian (% 18,4).

Ateratzen dugun lehen ondorioa da horrelako lan batek hiztegia aberasteko bidea eman dezakeela. Batez ere, kolokazioen kategorian da ondorio hori nabarmena, hiztegian daudenak % 15,3 baino ez baitira.

Esapide idiomatikoak dira hiztegietan proportzionalki errepresentatuen daudenak (% 28,75). Lehen adierazi dugu gehienak esapide figuratiboak direla. Sailkatzaileen artean erabateko adostasunik ez denerako ezarri ditugun irizpideak aplikatuz, esapide idiomatiko opakoak 8 lirateke, eta figuratiboak 72. Bada, opakoetatik % 75 daude hiztegietan (6), baina figuratiboen % 23,6 besterik ez (17). Populazio txikiak izan arren, proportzio horiek iradokitzen digute gaur egungo kazetaritza-testuetan badirela hiztegietan bildu ez diren hainbat esapide idiomatiko, gehienak figuratiboak, eta, seguru aski, berri samarrak euskararen erabileran.

Horrek guztiak zerikusia du, dudarik gabe, hizkuntza normalizazio-bidean egotearekin.

kategoria	Hiztegi-erreferentzia bai	ez	Totala
`id`	23	57	80
`col`	41	227	268
`free`	10	787	797
Totala	74	1 071	1 145

VI.5 Taula: Eskuz sailkatutako ebaluazio-erreferentziako bigramen eta hiztegi-erreferentziaren arteko konparazioa.

Erreferentzia sailkatuko bigramak aztertuz, ikusi da horietarik 17 II.4 atalean zenbait predikatu konplexuren inguruan egindako gogoetarekin erlazionatuta daudela. Batzuk (esaterako, *jaramon egin, mesede egin, muzin egin*) [XS+Adi] egiturakoak dira, non Adi aditz arina den; eta beste batzuk (hala nola *falta izan, espero izan, zor izan*) [XS+LOT/LAG] multzokoak. Horrelakoak esapide erdiidiomatikoetan sailkatu izanak II.2.2.3 atalean azaldu genuen Mel'čuk-en ereduak duen ikuspegi semantikoaren eragina salatzen du, eta ez dator erabat bat euskarazko fraseologian horrelakoak lokuziotzat jotzeko tradizioarekin. Horrenbestez, uste dugu komenigarria dela, kontrasterako, 17 kasu horiek esapide idiomatikotzat sailkatuz esperimentuen emaitzetan igarriko litzatekeen eragina aztertzea (ikus VII.3 atala).

VI.5 Ebaluazio-metrika

VI.5.1 Ranking-ataza

Tekniken emaitzak konparatzeko, neurri bakoitzak sortutako rankingetan oinarritu dugu ebaluazioa. Neurri "ideal" bat bagenu, emaitza bigramen kategorien multzo ordenatua izango litzateke, non `id` kategoria duten guztiak hasieran leudekeen, gero `col` kategoriakoak, eta azkenik, `free` kategoriakoak. Beraz, ideia da nolabait konparatzea idealki ordenatutako rankinga neurri bakoitzak sortutako rankingarekin. Horretarako, Spearman ρ edo Kendall τ heinen korrelazio-koefizienteak (*rank-correlation coefficient*) erabil litezke. Bigarrena aukeratu dugu guk. Arrazoi nagusia da koefiziente horren τ_B aldaerak berdinketen kudeaketa egokia egiteko aukera ematen duela (Gilpin, 1993). Kontuan izan behar dugu ordena idealtzat dugun rankingean berdin-

keta asko ditugula, `id`, `col` eta `free` klaseak bakarrik baitaude (1, 2 eta 3 heinak baino ez, beraz).

τ_B koefizientea τ koefizientearen moldaketa bat da, berdinketak dituzten heinen arteko korrelazioak kalkulatzeko. Honela definitzen da (Kendall, 1945; Agresti, 2010: 188-189)

$$\tau_B = \frac{C - D}{\sqrt{(n_o - T_X)(n_o - T_Y)}} \tag{VI.1}$$

Non:
C = bat datozen bikoteen kopurua (*concordant pairs*)
D = bat ez datozen bikoteen kopurua (*disconcordant pairs*)
$n_o = n(n-1)/2$ (bikote-kopurua, non n behaketa-kopurua den)
$T_X = \sum_i t_i(t_i - 1)/2$
$T_Y = \sum_j u_j(u_j - 1)/2$
t_i = lehen propietatearen i. heinaren barneko berdinketa-kopurua
u_j = bigarren propietatearen j. heinaren barneko berdinketa-kopurua

Esperimentalki lortutako rankingek oso berdinketa gutxi dituzte (edo bat ere ez), continuum bat osatzen duten neurketa-balioen ordenaziotik sortuak baitira. Beraz, aurreko ranking idealarekin konparatzean, ez da errealista goren lerrotzat (*topline*) 1 balioa kontsideratzea. Balio hori erdiesteko, ranking idealaren berdinketa guztiak birsortu beharko lirateke ranking esperimentalean. Neurri batek hautagaiak idealki ordenatuko balitu (`id` guztiak lehenik, gero `col` guztiak, eta azkenean `free` kategoriakoak), baina berdinketarik gabe, lortuko litzatekeen Kendall τ_B koefizientearen balioa 0,675 litzateke. Hori da, errealista izanda, lortzea espero dezakegun korrelazio maximoa.

Kendall τ_B koefizientea kalkulatzeko, Gene Boggsen Perl moduluak[7] ohiko Kendall τ koefizientea kalkulatzeko duen scripta modifikatu dugu, aurreko ekuazioaren araberako koefizientea kalkula dezan.

Horrez gain, batez besteko doitasunak ere kalkulatu dira (AP - Average Precision). AP neurria honela definitzen da (Zhu, 2004): egiazko positibo bat rankingeko k posizioan agertzen den bakoitzean P doitasunak hartzen dituen balioen batezbestekoa. P/R kurbaren azpiko azalera da. Praktikan, integral horren balioa honela adieraz daiteke:

$$AP = \sum_{k=1}^{n} P(k)\Delta r(k), \tag{VI.2}$$

[7] http://search.cpan.org/$\sim$gene/Statistics-RankCorrelation-0.1203/

non k rankingaren heina edo hautagai-kopurua den, n erauzitako hautagai guztien kopurua, $P(k)$ k heineko doitasuna, eta $\Delta r(k)$ $k-1$ heinetik k heinera pasatzean gertatzen den estaldura-aldaketa. Heineko hautagaia positibo faltsua denean, $\Delta r(k) = 0$, eta, beraz, heineko doitasuna ez da batezbestekoaren kalkuluan sartzen.

Froga daiteke honako hau aurrekoaren baliokidea dela (Turpin eta Scholer, 2006):

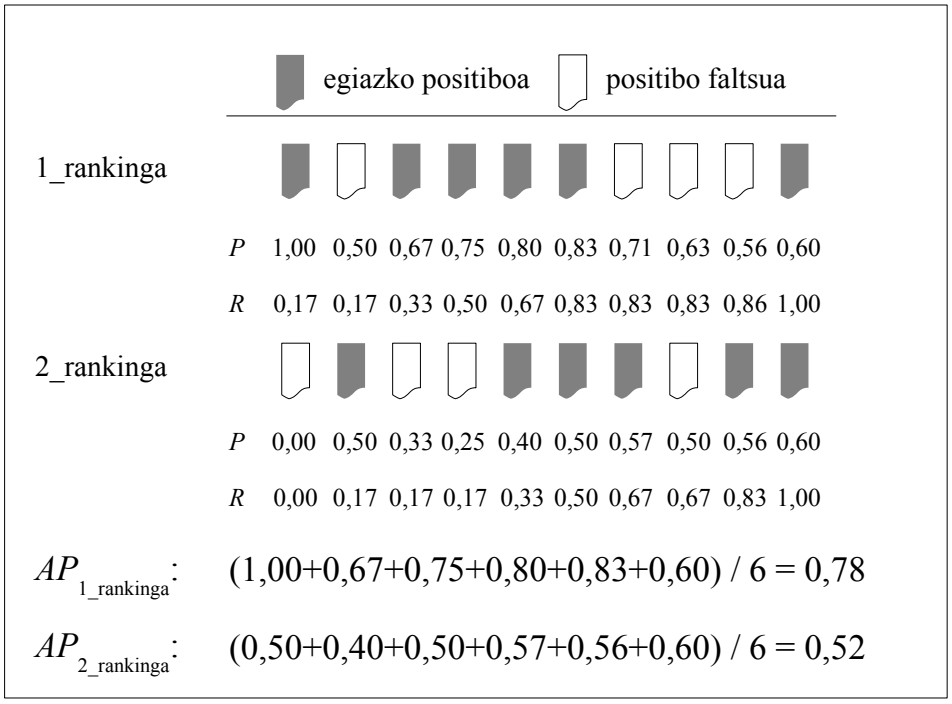
$$AP = \frac{\sum_{k=1}^{n}(P(k) \times \text{rel}(k))}{\text{zuzenak diren hautagaien kopurua}}, \tag{VI.3}$$

non $\text{rel}(k)$ adierazleak 1 balioa duen heineko hautagaia egiazko positiboa denean, eta 0, berriz, ez denean. VI.5 irudian, AP-ren kalkulua azaltzeko adibide bat eman dugu[8].

Hiru AP neurtu ditugu. Lehenik, $AP_{\texttt{UF}}$, `id` eta `col` bereizketa kontuan hartzen ez duena; bestetik, AP espezifikoak, esapide idiomatikoetarako ($AP_{\texttt{id}}$) eta kolokazioetarako ($AP_{\texttt{col}}$). Metrika hauek ez digute aukerarik ematen karakterizazio-esperimentu batek idiomatikotasun-continuuma zenbateko zehaztasunez sortzen duen ebaluatzeko, baina baliagarriak dira UFekiko eta horietako kategoria bakoitzarekiko duen portaera ikusteko, τ_B koefizientearen emaitzekin kontrastatzeko eta emaitzak hobeto interpretatzeko. $AP_{\texttt{id}}$ eta $AP_{\texttt{col}}$-en kasuan, kontuan hartu behar dugu bi horien balioak ezin direla aldi berean arbitrarioki handiak izan; nolabait ere, kontrajarrita baitaude. `id` kategoriako gehienak rankingaren goialdean badaude, `col` kategoriakoak horien ondoren agertuko dira, eta $AP_{\texttt{col}}$ txikiagoa izango da. Aldiz, kolokazioen erauzketan espezializatua den neurri batek `col` gehienak goiko posizioetan ordenatuz gero, ezinbestean izango da $AP_{\texttt{id}}$-a aurreko kasuan baino txikiagoa. Horren erakusgarri da ranking optimo errealista batek (zeinen τ_B lehen aipatutako 0,675 litzatekeen, eta $AP_{\texttt{id}} = 1$ lukeen), gehienaz ere, $AP_{\texttt{col}}$-en 0,563-ko balioa lor lezakeela. Hori guztia gogoan izan behar dugu esperimentuen emaitzak baloratzean.

Esperimentuek sortutako rankingetatik AP-ren balioak kalkulatzeko prozesua Perl lengoaian programatu da.

[8] Iturria: Rong Jin–Information Retrieval. Lecture 4. (http://www.cse.msu.edu/~cse484/evaluation.ppt)



VI.5 Irudia: Batez besteko doitasunaren (*AP – Average Precision*) kalkuluaren azalpena.

Esangura-testak

Kendall τ_B koefizientearen esangura-testak egin dira. Horretarako, p balioak kalkulatu dira, estatistikotzat z testa erabiliz, Bolboacă eta Jäntschiren (2006) lanaren arabera konputatuta[9].

$$Z_B = \frac{C - D}{\sqrt{v}} \tag{VI.4}$$

Non:
$v = (v_0 - v_t - v_n)/18 + v_1 + v_2$
$v_0 = n(n-1)(2n+5)$
$v_t = \sum_i t_i(t_i - 1)(2t_i + 5)$

[9] τ_B koefizientearen kalkulua azaltzean aipatu dugun Gene Boggsen Perl moduluak dakarren ohiko z testa egiteko kodea modifikatu dugu.

$$\upsilon_u = \sum_j u_j(u_j - 1)(2u_j + 5)$$
$$\upsilon_1 = \sum_i t_i(t_i - 1) \sum_j u_j(u_j - 1)/(2n(n - 1))$$
$$\upsilon_2 = \sum_i t_i(t_i - 1)(t_i - 2) \sum_j u_j(u_j - 1)(u_j - 2)/(9n(n - 1)(n - 2))$$

VI.5.2 Sailkapen-ataza

VII.2 atalean azalduko dugunez, sailkapen-atazari ikasketa automatikoaren bidez ekin diogu, eta Weka paketea erabili dugu esperimentu horiek egiteko (Hall et al., 2009).

Ikasketa automatikoko algoritmoak ebaluatzeko, Wekak eskaintzen dituen neurri hauek erabili ditugu:

- Zuzen sailkatutako instantzia-kopurua (*Correctly Classified Instances* - CCI). Zehaztasunaren baliokidea da (*accuracy*).
- Kategoria bakoitzeko F neurriak: $F_{\texttt{id}}$, $F_{\texttt{col}}$, F_{free}.
- Batez besteko F neurri haztatua (*Weighted Average F-measure*). Mikrobatezbestekoa da (F_{mikro}).

Mikrobatezbestekoaz gain, makrobatezbestekoa ere (F_{makro}) kalkulatu dugu.

Sailkatzailearen ebaluazioaren beste alderdi funtsezko bat zuzenean lotuta dago sailkatzaileak trebatzeko eta haren portaera ebaluatzeko erabiltzen den metodologiarekin. Hain zuzen ere, erabaki behar da eskura dugun erreferentzia sailkatutik zenbat adibide edo instantzia erabiliko diren sailkatzaileak ikas dezan (ikaste-multzoa) eta zenbat haren emaitzak ebaluatzeko (test-multzoa). Horretan kontuan hartu beharreko faktore garrantzitsu bat erreferentzian ditugun adibideen kopurua da, txikiegia denean aurreko bi multzoen ausazko bereizketa adierazgarria ez izatea gerta baitaiteke. Sailkatzailea trebatzeko eta ebaluatzeko metodologiaren zehaztapenak ikasketa automatikoko esperimentuen diseinuaren atalean emango ditugu (VII.2.1 atala).

VII. KAPITULUA

Idiomatikotasuna karakterizatzeko esperimentuak

Bi karakterizazio-ataza antolatu ditugu: ranking-ataza eta sailkatze-ataza. Lehenean, idiomatikotasunaren ezaugarriak edo propietateak modu independentean neurtu eta ebaluatu ditugu. Bigarrenean, neurketa horien emaitzak ikasketa automatikoan oinarritutako sailkatzaileak trebatzeko erabili ditugu, propietateen inguruko ezagutzak konbinatuz. Ataza horietan egindako lan esperimentala deskribatuko dugu hurrengo atal banatan.

VII.1 Propietateen banakako neurketa

Esperimentu bakunak egin ditugu idiomatikotasuna osatzen duten propietateak banaka neurtzeko. Horrelakoetan, propietate bakoitzaren neurketa (agerkidetza, konposizionaltasuna, malgutasun morfosintaktikoa eta malgutasun lexikala) saio bereizietan egin da, hainbat neurri estatistiko erabiliz. Bestela esanda, esperimentu bakoitzean propietate-mota bat eta neurri bat erabiltzen dira.

VII.1.1 Idiosinkrasia estatistikoaren neurketa, agerkidetza-informazioa darabilten elkartze-neurrien bidez

Bigramen agerkidetza-datuen analisi estatistikoa S. Everten UCS toolkit[1] paketearen bidez egin dugu (Evert, 2005). UCS toolkit agerkidetza-datuen analisi estatistikorako liburutegi- eta script-bilduma bat da. UCS/Perl azpisistemari esker, NSPk sortzen duen bigrama-fitxategiaren formatua inportatzeko aukera ematen du, eta oso eraginkorra da sorta zabal bateko AMak

[1] http://www.collocations.de/software.html

kalkulatzen eta testu-fitxategiari gehitzen. Gainera, fitxategi hori irizpide batzuen arabera ordenatzeko edo emaitzak iragazteko aukera ere ematen du.

VII.1 irudian, VI.3 atalean deskribatutako w1_f30_norm erauzketa UCSren bidez prozesatuz lortzen den informazioa dugu, t neurriaren arabera ordenatuta.

UCSk hainbat AM kalkulatzeko aukera ematen du[2]. Literaturan erabilienak diren elkartze-neurriak kalkulatu ditugu; zehazki, III.6.3 azpiatalean deskribatu ditugun hauek: z neurria, t neurria, khi karratua (χ^2), egiantz-arrazoiaren logaritmoa (LLR - *log-likelihood ratio*), Fisherren test zehatza, elkarrekiko informazioa (MI), MI^3 eta f.

VII.1.1.1 Emaitzak

VII.1 taulan bistaratu ditugu AMen araberako rankingen Kendall τ_B koefizientearen balioak (idiomatikotasun-ranking ideal batekiko hein-korrelazioak), eta batez besteko doitasunak, UFen rankingerako ($AP_{\texttt{UF}}$), eta AP espezifikoak esapide idiomatikoetarako eta kolokazioetarako ($AP_{\texttt{id}}$ eta $AP_{\texttt{col}}$). Everti jarraituz (Evert, 2005: 142), ausazko rankingaren araberako emaitzak eman ditugu oinarri-lerrotzat.

	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
ausazko rankinga	0,000	0,309	0,070	0,234
z neurria	(-0,038)	0,297	0,109	0,204
t neurria	**0,197**	**0,455**	0,084	**0,383**
χ^2	(-0,037)	0,302	**0,119**	0,206
LLR	0,156	0,427	0,100	0,335
Fisher	0,156	0,426	0,100	0,335
MI	(-0,121)	0,257	0,086	0,182
MI^3	0,103	0,389	0,107	0,291
f	0,189	0,436	0,074	0,379

VII.1 Taula: Elkartze-neurrien emaitzak, τ_B-ren balio ez-esanguratsuak parentesi artean ageri dira ($p > 0,05$).

Emaitza horietatik nabarmentzekoak diren alderdiak:

[2] http://www.collocations.de/UCS/UCS-Perl-html/UCS/AM.html

id	ll	l2	f	f1	f2	N	am.t.score	am.log.likelihood	am.MI	am.MI3	am.chi.squared	am.z.score
10443	nahi_nahi_IZE_ARR_ABS_MG	ukan_ADI	115928	143684	426774	2817230	276.5541924	333326.1246	0.72640428	10.85478097	505872.384	638.237604
2328	behar_behar_IZE_ARR_ABS_MG	ukan_ADI	102522	214368	426774	2817230	218.7697593	142397.6740	0.49928146	10.52091560	192738.624	388.712066
13257	uste_uste_IZE_ARR_ABS_MG	ukan_ADI	56044	56371	426774	2817230	200.6644221	214195.3541	0.81709774	10.31415599	317803.094	514.066433
2315	behar_behar_IZE_ARR_ABS_MG	izan_ADI	99512	214368	629138	2817230	163.6992368	66382.3628	0.31779189	10.31354280	77627.050	236.016626
11169	parte_parte_IZE_ARR_ABS_MG	hartu_ADI	18292	22636	93411	2817230	129.6985361	106445.8006	1.38688579	9.91140818	427467.972	640.291844
8823	kontuan_kontu_IZE_ARR_INE_NUMS	hartu_ADI	14494	19298	93411	2817230	115.0761443	79664.0841	1.35510024	9.67747675	312392.106	547.690769
226	ahal_ahal_IZE_ARR_ABS_MG	izan_ADI	22077	27128	629138	2817230	107.8104326	43202.0911	0.56159867	9.24947878	55065.450	205.806991
5239	ezin_ezin_IZE_ARR_ABS_MG	izan_ADI	26592	48587	629138	2817230	96.5327738	24447.4904	0.38930730	9.23880931	29920.507	151.122259
9137	lan_lan_IZE_ARR_ABS_MG	egin_ADI	18584	55657	300597	2817230	92.7606719	21256.9122	0.49545707	9.03373546	30750.560	164.093893
2365	behera_behe_IZE_ARR_ALA_NUMS	utzi_ADI	8278	16263	28987	2817230	89.1443575	55953.1793	1.69434379	9.53019464	399519.065	626.997424
3178	deia_dei_IZE_ARR_ABS_NUMS	egin_ADI	9791	11233	300597	2817230	86.8366814	35812.4581	0.91216895	8.89382306	69232.961	248.191952
6904	hitz_hitz_IZE_ARR_ABS_MG	egin_ADI	11337	20466	300597	2817230	85.9662667	24982.5302	0.71530284	8.82429913	43264.120	195.875215
4103	erabakia_erabaki_IZE_ARR_ABS_NUMS	hartu_ADI	7961	10511	93411	2817230	85.3184028	43424.5220	1.35874787	9.16068312	172623.824	407.771297
5248	ezin_ezin_IZE_ARR_ABS_MG	ukan_ADI	19090	48587	426774	2817230	84.8953307	17026.5353	0.41391020	8.97552205	22416.856	136.722157
9036	lagun_lagun_IZE_ARR_ABS_MG	hil_ADI	6423	17523	21605	2817230	78.4668589	41778.0415	1.67939749	9.29487334	298416.712	542.481604
4822	espero_espero_IZE_ARR_ABS_MG	ukan_ADI	9651	13323	426774	2817230	77.6951952	22108.4695	0.67959464	8.64873927	34180.878	169.899258
5308	falta_falta_IZE_ARR_ABS_MG	izan_ADI	10739	12829	629138	2817230	75.9830069	21979.0205	0.57384742	8.63577511	27991.057	147.109320
10516	neurriak_neurri_IZE_ARR_ABS_NUMP	hartu_ADI	6364	9912	93411	2817230	75.6549226	31080.8866	1.28699316	8.89445350	115038.258	332.915313
1664	aurrera_aurre_IZE_ARR_ALA_NUMS	egin_ADI	8739	17482	300597	2817230	73.5289580	17027.2496	0.67070829	8.55363176	28531.913	159.152166
9927	martxan_martxa_IZE_ARR_INE_NUMS	jarri_ADI	5316	7367	47148	2817230	71.2199219	35447.6728	1.63465333	9.08582328	223009.393	467.657775
2422	berri_berri_IZE_ARR_ABS_MG	eman_ADI	5480	10005	110195	2817230	68.7405445	22355.8524	1.14622388	8.62378500	69107.205	257.231949
11221	partida_partida_IZE_ARR_ABS_MG	jokatu_ADI	5036	14837	36012	2817230	68.2922156	25847.8516	1.42411469	8.82828613	126106.546	351.907519
341	akordioa_akordio_IZE_ARR_ABS_NUMS	lortu_ADI	4921	11638	33380	2817230	68.1841466	28691.4956	1.55251098	8.93661771	168597.863	407.322970
1279	asmoa_asmo_IZE_ARR_ABS_NUMS	ukan_ADI	7948	13798	426774	2817230	65.7058749	13209.5009	0.58006607	8.38058179	19442.330	128.125991
9103	lanean_lan_IZE_ARR_INE_NUMS	aritu_ADI	4232	13237	20797	2817230	63.5517367	25983.8462	1.63657770	8.88966902	177049.953	418.231052
13005	uko_uko_IZE_ARR_ABS_MG	egin_ADI	4990	5072	300597	2817230	62.9788289	21586.1728	0.96475896	8.36096005	41014.095	191.237942
5721	garaipena_garaipen_IZE_ARR_ABS_NUMS	lortu_ADI	3946	5860	33380	2817230	61.7118876	28121.9255	1.75459544	8.94690960	219489.561	465.228754
10014	merezi_merezi_IZE_ARR_ABS_MG	ukan_ADI	5221	5726	426774	2817230	60.2518198	16508.6343	0.77952670	8.21503409	25804.306	147.820109
1526	aukera_aukera_IZE_ARR_ABS_NUMS	eman_ADI	5918	33495	110195	2817230	59.8978544	9530.7166	0.65485539	8.19920532	17068.713	127.303095
441	alde_alde_IZE_ARR_ABS_MG	egin_ADI	7610	22716	300597	2817230	59.4508598	8598.2687	0.49689046	8.25965977	12523.484	105.342439
1592	aurka_aurka_IZE_ARR_ABS_MG	jokatu_ADI	3939	18423	36012	2817230	59.0091985	15998.7735	1.22340069	8.41417265	59384.999	241.334862
1537	aukera_aukera_IZE_ARR_ABS_NUMS	izan_ADI	14472	33495	629138	2817230	58.1212580	7294.6571	0.28662496	8.60768207	8516.205	80.843943
13517	zigorra_zigor_IZE_ARR_ABS_NUMS	ezarri_ADI	3435	9875	9045	2817230	58.0679195	28214.3354	2.03480337	9.10665685	367787.921	604.418782
4712	esku_esku_IZE_ARR_ABS_MG	egon_ADI	4105	11195	111036	2817230	57.1836079	12532.2669	0.96864756	8.19527388	31796.789	174.419619
8700	konpromisoa_konpromiso_IZE_ARR_ABS_NUMS	hartu_ADI	3607	5736	93411	2817230	56.8915695	17282.5407	1.27796128	8.39225357	63618.923	247.758496
2684	bilera_bilera_IZE_ARR_ABS_NUMS	egin_ADI	4540	7073	300597	2817230	56.1790051	11716.1082	0.77928986	8.09340156	21307.480	137.790404
3455	egia_egia_IZE_ARR_ABS_NUMS	esan_ADI	3247	10416	19692	2817230	55.7047580	19953.7428	1.64931379	8.67227837	139879.270	372.005353
1662	aurre_aurre_IZE_ARR_ABS_MG	egin_ADI	4889	9558	300597	2817230	55.3359508	9745.2555	0.68069067	8.05913075	16488.600	121.158172
6139	gogora_gogo_IZE_ARR_ALA_NUMS	ekarri_ADI	2995	2995	6659	2817230	54.5972376	37833.7311	2.62641329	9.57920694	1265448.039	1122.993686
8825	kontua_kontu_IZE_ARR_ABS_NUMS	izan_ADI	5282	5904	629138	2817230	54.5359707	12208.1738	0.60272844	8.04832523	15373.132	109.155913
7876	izena_izen_IZE_ARR_ABS_NUMS	eman_ADI	3684	10253	110195	2817230	54.0885616	11118.1233	0.96312908	8.09576832	28070.609	163.934353
13138	urte_urte_IZE_ARR_ABS_MG	eraman_ADI	3257	27452	18272	2817230	53.9503169	13696.4238	1.26227999	8.28791550	54118.602	230.745900
1099	artean_arte_IZE_ARR_INE_NUMS	egon_ADI	4317	19868	111036	2817230	53.7858469	8489.3981	0.74138639	8.01175049	16720.891	126.287745
77	adierazpenak_adierazpen_IZE_ARR_ABS_NUMP	egin_ADI	3702	4748	300597	2817230	52.5177096	11835.3120	0.86376337	8.00063620	22600.076	141.967018
7095	ihes_ihes_IZE_ARR_ABS_MG	egin_ADI	3670	4747	300597	2817230	52.2197114	11621.6551	0.86008450	7.98941662	22155.841	140.564844
12135	saria_sari_IZE_ARR_ABS_NUMS	jaso_ADI	2797	8029	20460	2817230	51.7841206	17634.5877	1.68094768	8.57433262	129940.920	358.649680
216	ahalegina_ahalegin_IZE_ARR_ABS_NUMS	egin_ADI	3285	3746	300597	2817230	50.3412330	12042.6906	0.91480526	7.94787601	23347.115	144.319987
3457	egia_egia_IZE_ARR_ABS_NUMS	izan_ADI	6247	10416	629138	2817230	49.6080625	6847.0453	0.42904692	8.02038993	8541.160	81.297226
5427	finala_final_IZE_ARR_ABS_NUMS	jokatu_ADI	2557	5128	36012	2817230	49.2704813	15431.6120	1.59115777	8.40661923	96097.014	307.727063
774	apustua_apustu_IZE_ARR_ABS_NUMS	egin_ADI	3125	3590	300597	2817230	49.0494719	11351.1259	0.91159324	7.90129328	21999.768	140.097683
9061	laguntza_laguntza_IZE_ARR_ABS_NUMS	eman_ADI	3201	10912	110195	2817230	49.0333906	8233.2723	0.87504173	7.88561308	18838.202	134.280451
4495	erresuma_erresuma_IZE_ARR_ABS_0	batu_ADI	2388	2388	4257	2817230	48.7933257	32716.8549	2.82071865	9.57678730	1579299.573	1255.219703
13325	zalantzan_zalantza_IZE_ARR_INE_NUMS	jarri_ADI	2447	2688	47148	2817230	48.5577644	18530.1901	1.73556373	8.51283167	130564.573	358.129875

VII.1 Irudia: UCS toolkit-ek sortzen duen agerkidetza-informazioa, `ucs-sort` komandoa erabiliz t neurriaren arabera ordenatuta.

- Oro har, AMek kolokazioen rankingean lortzen dituzte emaitza nabariak, eta esapide idiomatikoen kasuan oinarri-lerroaren mailako portaera dute.

- t neurria da `id/col/free` ranking idealera gehien gerturatzen den neurria, baina f-rekiko aldeak ez dira esanguratsuak. f-ren emaitzak onenen artean egotea bat dator zenbait lanetan lortutako emaitzekin (Krenn eta Evert, 2001; Wermter eta Hahn, 2006; Frontini et al., 2012).

- z neurriaren, MIren eta χ^2-ren τ_B balioak ausazko oinarri-lerroaren azpitik daude. Esapide idiomatikoen AP-ren balio onenak, berriz, χ^2−k lortu ditu, eta z neurria da bigarrena; dena den, ez dute askogatik gainditzen oinarri-lerroa.

- AP-ren balioen kasuan, kolokazioekin emaitza onenak dituzten AMak (t neurria eta f), azkenak dira esapide idiomatikoen emaitzetan. Espero zitekeen emaitza da hori, VI.5.1 atalean $AP_{\texttt{id}}$ eta $AP_{\texttt{col}}$-en balioen arteko erlazio kontrajarriaz esandakoa kontuan izanik.

- Antzeko erdi-mailako profila agertzen dute Fisherren test zehatzak eta LLR neurriak. Parekotasun horrek baieztatzen du III.6.3 atalean aurreratu genuen Everten iritzia (Evert, 2005: 112).

AMetan UF-kategoriaren arabera hauteman ditugun portaera-aldaketak aski bat datoz Evertek aipatutakoekin (Evert, 2005: 146). Gure esperimentuetan bezala, harenetan ere χ^2 eta PMIren emaitzak, bere txikian, hobeak dira esapide idiomatikoen kasuan kolokazioetan baino. Alderantzizkoa gertatzen da f eta t neurriarekin.

Kontuan izanik agerkidetzan oinarritutako teknikak direla UFak, bereziki kolokazioak, eraузteko ohiko prozedura, kontsidera genezake AMen emaitza horiek direla, nolabait, gure oinarri-lerroa. Idiomatikotasunaren gainerako propietateak neurtuz horiek baino emaitza esanguratsuki hobeak lor daitezkeen egiaztatzea da, hain zuzen ere, ikerketa honen egiteko nagusia.

VII.1.2 Konposizionaltasun semantikoaren neurketa, antzekotasun distribuzionalaren bidez

VII.1.2.1 Metodologiaren oinarriak

III.7 atalean deskribatu ditugun tekniketan inspiratuta, hauek dira UFen konposizionaltasuna karakterizatzeko gure metodologiaren oinarriak:

- UF izateko hautagai den hitz-konbinazioaren testuinguruak haren osagai bakunen testuinguruekin konparatzea. Konparazio hori osagai bakoitzarekiko egin dugu, osagai bakoitzaren eragina bereiz aztertzeko, eta, horrez gain, konposizionaltasunaren baterako neurria ere kalkulatu da, bien batezbestekoaren bidez.

- Osagai bakunen testuinguruetan konbinazioaren testuinguruak ez sartzea. Esaterako, *mahaia jaso* bigramaren agerpena atzematen denean, testuinguruko hitzek *mahaia jaso*ren testuinguru-dokumentua elikatu dute, eta ez dira *mahai* eta *jaso*ren testuingurutzat kontsideratu. Beraz, *mahai*ren testuinguruetatik *mahaia jaso* bigramari dagozkionak kendu dira (*jaso* ager liteke *mahai*ren testuinguruetan, baina agerkidetza deklaratzeko ezartzen den baino distantzia handiagoko agerpenetatik letorke, hau da, bigrama-mailan erlazionatu gabeko agerkidetzatik). Horrela jokatuz, ahal den gehiena bereizten dugu konbinazioaren eragina osagaien testuinguruen modelizazioan, eta konparazioa adierazgarriagoa izan daiteke.

Antzekotasun distribuzionaleko tekniken bidez egin dugu testuinguruen arteko konparazioa. Zehazki, III.7 atalean deskribatutako neurri eta teknika hauek aplikatu ditugu:

- WSM (Word Space Model) ereduan ohikoak diren eta III.7.2 azpiatalean azaldu ditugun antzekotasun-neurrietatik, hauek aukeratu ditugu: Jaccard koefizientea, kosinua eta Jensen-Shannon dibergentzia.

- Berry-Roggheren (1974) R balioa (R_{BR}) eta Wulffek (2008) egindako bi hedapenak (R_{W_1} and R_{W_2}).

  Nahiz eta WSMren ohiko antzekotasun-neurrien artean ez diren aipatzen, kidekoak direla esan daiteke. Izan ere, III.7.3 azpiatalean adierazi bezala, informazio bera behar da horiek kalkulatzeko eta Jaccard koefizientea kalkulatzeko (bektoreen balioak edo pisuak ez dira kontuan hartzen).

- VSMren inplementazio berezia den LSA (*Latent Semantic Analysis*) teknikaren aplikazioa esploratu dugu.

Bestetik, UFen konposizionaltasuna neurtzeko argitaratu diren ikerkuntzetan IR sistemez baliatzen den lanik aurkitu ez badugu ere, komenigarritzat jo dugu dokumentu arteko antza kalkulatzeko indize batzuk ere aplikatzea. Ataza desberdina izan arren, IR sistema askotan VSM eredua da

oinarria, beste zenbait metodologiarekin batera (Otegi, 2012: 23), eta interesgarria iruditu zaigu horiekin esperimentatzea. Ideia gakoa da testuinguruak dokumentutzat tratatzea.

Azkenik, adiera-indukzioaren teknika hiponimia-hiperonimia kasuetara aplikatu da, eta, Korkontzelos eta Manandharrek (2009) erakutsi bezala, baliagarria gertatu da ingelesezko `noun+noun` eta `adjetive+noun` konbinazioekin. Zalantzazkoa da, ordea, hiponimia-hiperonimia hurbilpena egokia den `izena+aditza` konbinazioetarako; hori dela eta, teknika hau ez dugu gure ikerkuntzan aplikatu.

Testuinguruaren errepresentazioaren aldetik, geuk eratutako bektore-espazioa osatzeko eta IR esperimentuetarako testuinguru-dokumentuak osatzeko behar den lehen urratsa berbera da:

- Konparatu behar diren bigrama hautagaien testuinguruak eta bakoitzari dagozkion unigramen testuinguruak dokumentu banatan sortzea (aztergai den bigrama bakoitzeko, hiru dokumentu).

Ondoren, teknika bakoitzaren prozedura eta formatu-ezaugarrien arabera, berariazko prozesamendua aplikatu behar da:

- Geuk eratutako bektore-espazioaren kasuan, dokumentuetatik bektoreak sortzeko eta neurri horien bidez konparatzeko programazioa egin dugu (Perl).

- IR erako esperimentuetarako, Lemur Toolkit[3] (Allan et al., 2003) aukeratu dugu testuinguru-dokumentuen arteko antza kalkulatzeko. Lemurrek dakartzan indize hauek erabili ditugu: Indri, *tf-idf*, KL dibergentzia eta Okapi. Gainera, Lemurrek dakarren kosinuaren inplementazioa ere erabili dugu[4]. VII.1.2.3 azpiatalean zehaztuko dugu nola inplementatu dugun testuinguru-dokumentuen arteko konparazioa.

LSA bidezko esperimentuak direla eta, komeni da analizatzea LSA nola erabil daitekeen hitz-konbinazioen eta haien osagaien arteko antzekotasun distribuzionala neurtzeko. Izan ere, konbinazioen eta haien osagaien arteko testuinguru-antzekotasuna neurtzeko, eskura dugun lehen aukera da LSA eredua eratzeko aplikazioak prozesatu behar duen corpusean konbinazioak unitate gisa tratatzea, matrizearen errenkadetan ager daitezen, eta osagaiei dagozkien errenkadetako bektoreekin konparatu ahal izan ditzagun. Hori

[3] http://www.lemurproject.org
[4] http://www.lemurproject.org/doxygen/lemur/html/RetEval.html

bide da Baldwin et al.-en (2003) eta Katz eta Giesbrechten (2006) lanetan erabilitako prozedura, baina badu desabantaila bat, berriki Krcmár et al.-ek (2013) argi adierazi dutenez:

> ≪Treating the expressions as the single words affects the WSM vectors of their constituents. As an example, consider the replacement of occurrences of *short distance* by e.g. the EXP#123 label. This affects the WSM vectors of *short* and *distance* since the numbers of their occurrences and the numbers of contexts they occur in drops. Consequently, this also affects the methods for determining the compositionality which are based upon using the vectors of expressions' constituents.≫

Hori konpontzeko, hau da Krcmár et al.-ek (2013) proposatutako irtenbidea:

> ≪We added vectors for the examined expressions to WSM in such a way that the original vectors for words were preserved.≫

Hala ere, prozedura horren bidez ez da lortzen gure metodologiaren bigarren oinarrian ezarri dugun helburua, alegia, "osagai bakunen testuinguruetan konbinazioaren testuinguruak ez sartzea". Izan ere, osagaien jatorrizko testuinguruak atxikitzen badira, ez dira aztergai den konbinazioaren araberakoak. Hau da, lehen aipatutako *mahaia jaso*ren kasuan, *mahai* unigramaren testuinguruei eragiten die beste bigrama baten osagaia izateak, zeren, esaterako, *mahaia batu* bigramari esleitu zaizkien testuinguruak kontuan hartzekoak bailirateke *mahaia jaso*ren konposizionaltasuna neurtzeko behar diren *mahai*ren testuinguruetan.

Irtenbide bat izan liteke bigrama bakoitzeko LSA eredu bat eratzea, neurketa independenteak egiteko; posible izanda ere, ez dirudi bideragarria denik gure agertokian (hiztegigintzan). Eginbeharrari ekiteko beste modu bat izan daiteke, esaterako, dokumentuak konparatzeko modalitatean erabiltzea LSA. Corpus bakarra prozesatu beharrean, sistemak dokumentu-multzo baten LSA eredua sortzen du, non dokumentu horiek bigramen eta haietako bakoitzaren osagaien testuinguru-dokumentuak diren (lehen deskribatu dugun eran sortutakoak). Antzekotasuna dokumentuen artean egiten da, ez hitz- edo hitz-konbinazioen artean. Infomap softwareak[5] ematen du horretarako aukera, eta hori izan da hobetsi dugun bidea. Nolanahi ere, zenbait esperimentu egin ditugu konbinazioak unitate gisa adierazita dauden corpus batekin, besterik ez bada, konparazio gisa ebaluatzeko.

[5] http://infomap-nlp.sourceforge.net/

VII.1.2.2 Testuinguru-dokumentuen sorkuntza

Ebaluazio-lagineko bigrama bakoitzetik, hiru testuinguru-dokumentu sortu ditugu:

- Bigramaren testuinguruak.
- Bigramako izenaren testuinguruak, non bigramako aditza ez den agertzen.
- Bigramako aditzaren testuinguruak, non bigramako izena ez den agertzen.

Testuinguruak sortzeko, zenbait parametro zehaztu behar dira:

- Testuinguruetan sartuko diren hitz-motak: kategoria guztietako hitzak ala kategoria jakin batzuetakoak.
- Testuinguruetan hitzen formak ala lemak sartuko diren. Beste aukera bat da, izenetarako, bigramak normalizatzeko erabili dugun `lema_kasua` gakoa.
- Uneko hitz edo konbinazioaren eskuinera eta ezkerrera dauden testuinguruetatik zenbat hartuko den (leiho jakin baten barnekoak, esaldi osoa, paragrafo osoa...).
- Bigramaren osagaiak bata bestetik zer distantziatara agertu behar duten esaldian, bigrama-agerkidetzatzat hartzeko.

Gure esperimentuetan, aukera hauek egin ditugu aurreko kasu bakoitzerako:

- Testuinguru-dokumentuak elikatzeko, eduki-hitzak bakarrik erabiltzea, izen, aditz eta adjektibo kategorietakoak.
- Izenen kasuan, `lema_kasua` erabiltzea; gainerakoetan, lema. Gogoan izan behar dugu `izena+aditza` bigramen forma kanonikoan izenaren kasua barneratu dugula (IV.2 atala), *kontu hartu* / *kontuan hartu*, edo *egunkaria irakurri* / *egunkarian irakurri* bereizi ahal izateko.
- Esaldi osoa hartu dugu testuingurutzat, Reddy et al.-en (2011) lanari jarraituz.
- Ondoz ondoko agerkidetzak hartu dira bigramatzat; arrazoia da erauzketan ezarritako irizpide bera erabiltzea ($w = \pm 1$).

Lan horiek denak Perl lengoaian programatu dira.

VII.1.2.3 Testuinguruen prozesamendua

Testuinguru-dokumentuak sortu ondoren, hiru eratara prozesatu ditugu:

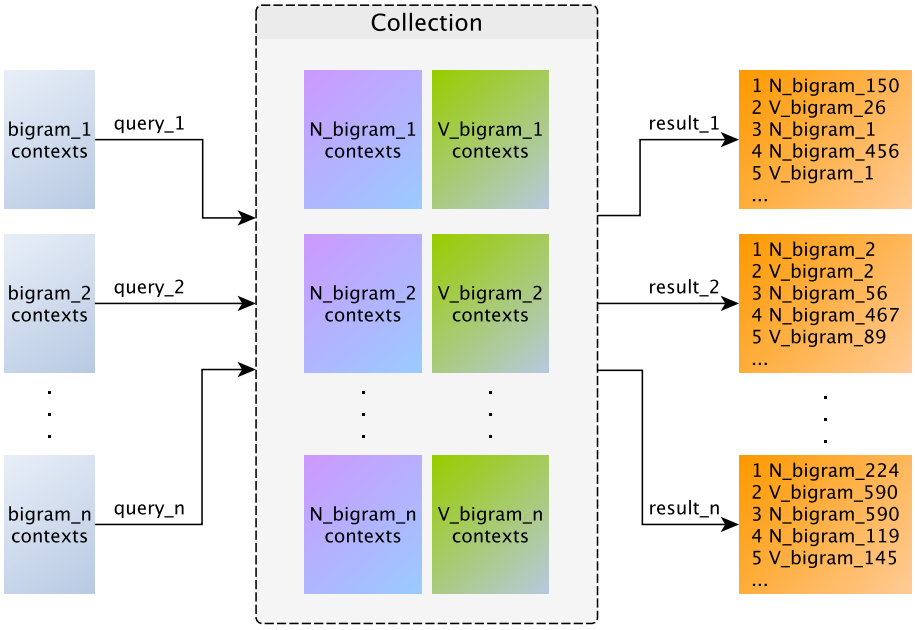
- Jaccard koefizientearen, kosinuaren, Jensen-Shannon dibergentziaren, R balioaren eta haren hedapenen kasuan, bektoreak eratu dira eta WSM eredua eraiki. Parametro desberdin hauen araberako esperimentuak egin ditugu:
  - Bektore desberdinak sortu ditugu, pisuetan edo balioetan neurri estatistiko hauek erabiliz: f, t neurria, LLR, PMI eta Fisherren test zehatza.
  - Antzekotasuna kalkulatzean, bektore osoaz gain, bektorea mugatzeko proba desberdinak egin ditugu dimentsionaltasuna gutxitzeko:
    * Ehuneko jakin bat: rankingaren % 75 eta % 50.
    * Kopuru mugatu bat: rankingeko lehen 3 000, 2 000 eta 1 000 elementuak. Reddy et al. (2011) eta Mitchell eta Lapata (2008) lanetan, 2 000 elementu erabili dira.
    * Maiztasun-atari bat ezarriz ($f > 3$), hortik beherakoak kontuan ez hartzeko.

    Bektorearen parte horiek hartzeko, elementuen pisuen araberako ranking desberdinak erabili ditugu.
  - Corpuseko maiztasun handieneko 100 hitzak ez dira kontuan hartu (*stop word* gisa prozesatu dira).

  Lan horiek denak Perl lengoaian programatu dira.
- Lemur: ideia nagusia da bigramen testuinguruen dokumentuak *query* edo kontsulta gisa erabiltzea bigramen osagaien testuinguruen dokumentuek osatzen duten bildumaren kontra, uneko bigramaren testuinguru-dokumentuen antzekoenak diren testuinguru-dokumentuak lortzeko. Ideia hori bi eratara inplementatu dugu:
  - L1: bektoreekin bezala, bigrama baten testuinguruak dokumentu bakarrean bildu dira, eta orobat osagaien testuinguruak. Horretara, bigrama bakoitzetik hiru dokumentu eratu dira: bigramaren testuinguruak dituena, kontsulta gisa erabilia, eta osagai bakoitzaren testuinguruak dituztenak, Lemurrek indizea sortzeko erabiltzen duen bilduman sartzen direnak.

– L2: bigramen testuinguru-esaldiak dokumentu banatan sartu dira; beraz, bigrama bakoitzerako agerkidetza adina kontsulta-dokumentu eratu dira. Osagaien dokumentuak, berriz, L1 esperimentuan bezala tratatu dira.

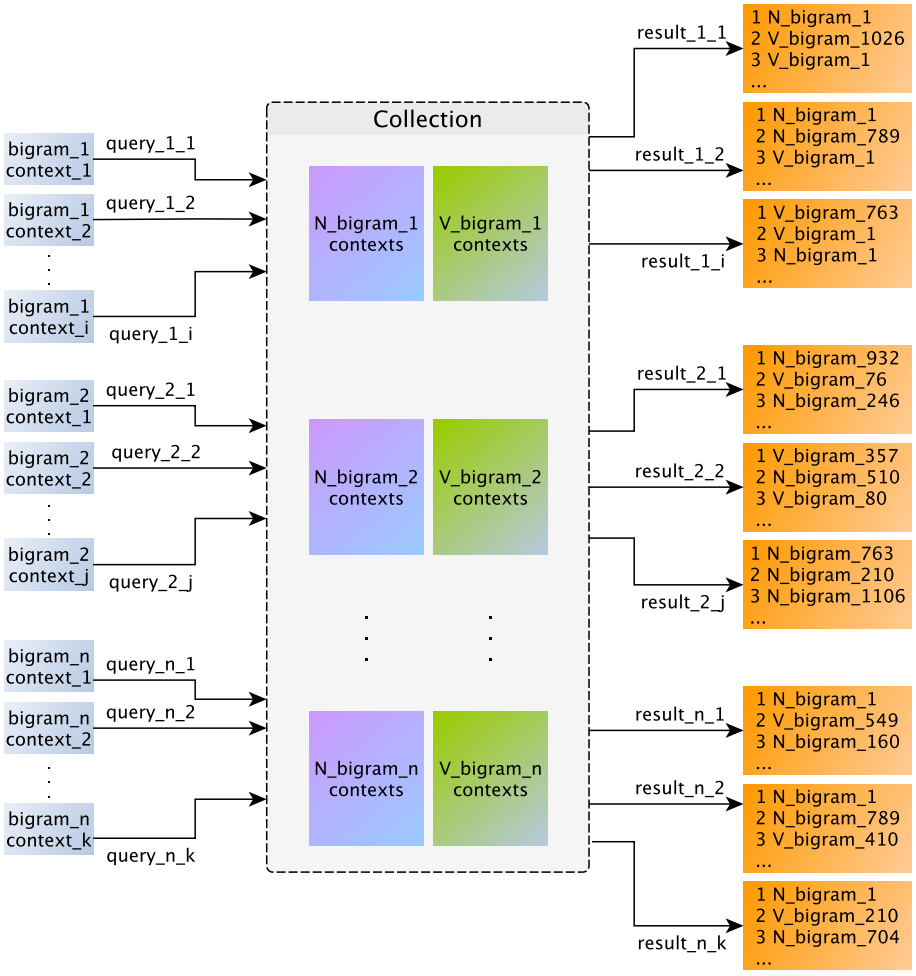
Beste modalitate batekin ere egin ditugu atariko proba batzuk: bigramaren eta osagaien testuinguruetako esaldi bakoitzak bere dokumentua izatea. Beraz, bigrama bakoitzerako agerkidetza adina kontsulta-dokumentu eratzen dira; irizpide bera erabiliz, osagai bakoitzaren esaldi bakoitza bildumako dokumentu bat da. Ebaluazio-laginaren erdiarekin egindako esperimentuek erakutsi zuten modalitate honen emaitzak aurrekoenak baino askoz apalagoak zirela (Gurrutxaga eta Alegria, 2012), eta ez dugu bide horretatik aurrera egin.



VII.2 Irudia: Lemurrekin egindako L1 modalitateko kontsultak eta emaitzak.

VII.2 irudian dugu L1 modalitatearen eskema, eta VII.3 irudian, L2 modalitatearena.

Diseinu hori inplementatzean, prozesatze-arrazoiak direla medio, Lemurren 2 000ko maximo batera mugatu ditugu testuinguru-esaldiak,



VII.3 Irudia: Lemurrekin egindako L2 modalitateko kontsultak eta emaitzak.

testuinguru guztietatik ausaz aukeratuta. WSMko bektore-sorkuntzan ere irizpide bera erabiliz egin ditugu probak, emaitzak albait konparagarrien egitearren.

Azkenik, Lemurren parametro-esleipen lehenetsia erabili dugu, eta kontsulta bakoitzak berreskuratzen dituen antzeko dokumentuen kopurua esperimentu-modalitatearen araberakoa izatea erabaki dugu:

– L1 modalitatean, unigramen dokumentu-kopuru bera da, hau da,

2 290 (1 145 x 2). Horretara, bigramei dagozkien unigramen dokumentuak antzeko dokumentuen rankingean zein posizio zehatzetan dauden jakin dezakegu, eta bigramen arteko antzekotasun-ranking zehatzak osatu.

– L2 modalitatean, berriz, berreskuratzen den antzeko dokumentuen kopurua txikiagoa da (500). Horren arrazoia bigramen rankingen sorkuntzaz jardutean azalduko dugu xeheago, hurrengo atalean.

Hauek dira *abenturak kontatu* bigramaren L1 modalitateko esperimentuen Indri indizearen araberako emaitzaren lehen 10 lerroak:

```
1 Q0 bizipenakEZ_kontatu 1 -8.4171 Exp
1 Q0 abenturak_kontatuEZ 2 -8.7138 Exp
1 Q0 bizitzaEZ_kontatu 3 -8.74488 Exp
1 Q0 abenturakEZ_kontatu 4 -8.90255 Exp
1 Q0 gezurtzat_joEZ 5 -8.92217 Exp
1 Q0 bizkarrezurrari_eutsiEZ 6 -8.93283 Exp
1 Q0 galbahetik_pasatuEZ 7 -8.94522 Exp
1 Q0 maisutasuna_erakutsiEZ 8 -8.94631 Exp
1 Q0 lurpetik_ateraEZ 9 -8.9499 Exp
1 Q0 sekretupetik_ateraEZ 10 -8.95369 Exp
```

- LSA: Infomapen kasuan, Lemurrekin erabilitako L1 modalitatearen antzeko prozesua egin dugu. Zehazki, aurreko esperimentuetako testuinguru-dokumentuetatik abiatu gara, eta Infomaperako behar den formatua eman diegu dokumentuei. Ondoren, Infomapen "`many` corpus" aukera erabili dugu LSA eredua eratzeko. Matrizeak 20 000 errenkada ditu (maiztasun handieneko 20 000 hitzak) eta horien bektoreek 1 000 dimentsio dituzte (matrizearen zutabeak), esanahia errepresentatzeko erabiltzen diren maiztasun handieneko eduki-hitzak. Matrizearen dimentsioak murrizteko, SVDrako baliotzat, 100 erabili da (hau da, 1 000 dimentsiotatik 100 balio singularretara murriztu dugu matrizea). `associate` komandoaren bidez, dokumentu baten antzekoenen rankinga lor dezakegu. Infomapek kosinua erabiltzen du bektoreen antzekotasun-neurritzat, eta, instalazio lehenetsian, 200 da eskura daitekeen ranking-luzera handiena.

  Hona hemen *abenturak kontatu* bigramaren lehen 10 dokumentu antzekoenen rankinga:

```
1: abenturak_kontatu: 1.000000
2: bizitza_kontatu: 0.951297
3: bizipenak_kontatu: 0.930322
4: abenturak_kontatuEZ: 0.928679
5: bizipenakEZ_kontatu: 0.909546
6: abenturakEZ_kontatu: 0.906421
7: bizitzaEZ_kontatu: 0.904043
8: zinemagilea_jaioEZ: 0.896375
9: munduan_barneratu: 0.892924
10: bizipenak_kontatuEZ: 0.866979
```

VII.1.2.4 Rankingen sorkuntza

WSMra aplikatutako antzekotasun-neurrien kasuan, ebaluazio-erreferentziako 1 145 bigramen rankingak antzekotasun-neurriaren balioen goranzko ordenaren arabera sortzen dira. Antzekotasun-neurria zenbat eta txikiagoa den, hainbat eta gorago agertu behar du bigramak ez-konposizionaltasunaren rankingean.

Lemurrekin egindako esperimentuetan, berriz, bigramen rankingak sortzeko, bigrama-osagai bakoitzaren dokumentuak dagokion bigrama-kontsultaren emaitzetan duen posizioa erabiltzen da. Indriren kasuan behintzat, Lemurren garatzaileek berek aitortzen dute kontsulta batean ematen diren *score* edo neurriak ez direla kontsulten artean erabiltzeko adierazgarriak[6]. Aurreko ataleko *abenturak kontatu* bigramaren L1 modalitateko esperimentuaren dokumentu-rankingetik, interesatzen zaigu `abenturak_kontatuEZ` unigrama 2. posizioan dagoela *abenturak kontatu* bigramarekiko antzekotasun-rankingean, eta `abenturakEZ_kontatu`, berriz, 4.ean. *adarra jo* bigramaren kasuan, 1330. posizioan agertzen da `adarra_joEZ` unigrama, eta `adarraEZ_jo`, 786.ean.

Bigrama guztien horrelako informazioa erabiliz, izenarekiko eta aditzarekiko antzekotasunaren araberako alderantzizko bigrama-ranking bana osatu dugu (posizioen beheranzko ordenan, hau da, antzekotasunaren kontrako ordenan edo ez-konposizionaltasunaren arabera). Bigramen ez-konposizionaltasunaren ranking bateratua egiteko, Lemurrek ematen dituen izenarekiko eta aditzarekiko posizioen batura egin dugu, eta horien alderantzizko ran-

[6] "... it's probably best to assume that these values are not comparable across queries." (http://sourceforge.net/p/lemur/wiki/Indri%20Document%20Scoring/)

kinga eratu. VII.2 taulako emaitza lortzen dugu erreferentziako aurreko 10 bigrametarako.

id	bigrama	Indri_rankN	Indri_rankV	Indri_rankNV
15	abenturak_kontatu	767	1041	1090
36	abiadura_moteldu	905	659	770
39	abiapuntura_itzuli	767	303	369
40	abiapuntua_izan	295	594	615
41	abilezia_erakutsi	375	481	532
63	adarretatik_heldu	905	98	138
65	adarra_jo	9	77	6
69	adibide_izan	108	485	390
101	aditzerat_eman	905	164	205
107	administrazioa_euskaldundu	618	1072	1075

VII.2 Taula: Ebaluazio-erreferentziako lehen 10 bigrametarako L1 esperimentuen emaitzetatik sortutako ranking-posizioak.

L2 esperimentuetan, zeinetan bigramaren testuinguru-esaldiak dokumentutan banatu baitira, bi prozedura erabili ditugu. Aurrez, gogoan izan behar dugu modalitate honetan ez dugula bildumako dokumentu guztiekiko rankinga eskuratzen, 500era mugatu baitugu itzultzen diren dokumentuen kopurua. Arrazoi bat da emaitzen bolumena arrazoizko kopuruetara ekartzea; beste bat, jarraian ikusiko dugunez, emaitzak prozesatzeko moduarekin dago erlazionatua.

Lehen aukera da bigrama baten N kontsulten emaitzetan dagokion unigramaren dokumentua zenbat aldiz itzultzen den kontuan hartzea (edo *hit*-kopurua). Kopuru hori bigramaren dokumentu-kopuru totalarekiko haztatuz, `hit_erl` lortzen dugu:

$$\texttt{hit\_erl} = \frac{\sum_{i=1}^{i=N} hit_i}{N} \tag{VII.1}$$

non N bigramaren dokumentu-kopurua den, eta hit_i funtzio bat, 1 balioa duena bigramarekin konparatzen ari garen osagaiaren dokumentua rankingeko zatian agertzen denean, eta bestela, 0. Bigrama baten dokumentuen lehen 500 antzekoenetan unigramaren dokumentua beti agertzen bada, `hit_erl` 1 da, eta, behin ere agertzen ez bada, 0.

Neurri horrek ez du kontuan hartzen unigramaren dokumentua rankingaren zein posiziotan gertatzen den. Zentzuzkoa dirudi datu hori kontuan hartzeak, eta, horretarako, bigramaren kontsulten emaitzetan unigramari dagozkion dokumentuak agertzen diren posizioen batezbestekoa kalkulatu eta haztatu egiten da, bigramak duen dokumentu-kopurua kontuan hartuz. Kalkulu horretan, alderantzizko ranking-posizioak erabiltzen ditugu.

Bigrama erabat konposizionala balitz, kontsulta guztietan 1. posizioan agertuko litzateke (beraz, alderantzizko rankingean, 500). Bigrama horrek N testuinguru baditu, rankingen batura $N\times$ 500 litzateke. Hori da bigramak lor lezakeen emaitza handiena, normalizaziorako erabiliko duguna, ranking haztatuaren (`rank_hazt`) 0-1 bitarteko eskala bat osatzeko. Erabat konposizionala izatera, `rank_hazt`-ek 1 balioa du.

Bigramaren dokumentu bati dagokion bildumako unigramaren dokumentua rankingean agertzen ez denean (hau da, 500. posiziotik behera dagoenean), alderantzizko rankingean 0 balioa esleitzen zaio. Erabat idiomatikoa bada, hau da, bigrama baten kontsulten emaitzetan unigramaren bildumako dokumentuak lehen 500en artean behin ere agertzen ez badira, `rank_hazt` 0 da. Beraz:

$$\texttt{rank\_hazt} = \frac{\sum_{i=1}^{i=N} r_i}{N \times 500}, \tag{VII.2}$$

non N bigramaren dokumentu-kopurua den, eta r_i, bigramaren dokumentu bakoitzaren kontsultak itzultzen duen rankingean osagaiaren dokumentuak duen posizioa.

Infomapen kasuan, antzeko metodoa erabili dugu, baina, kasu honetan, bigramaren osagaien dokumentuek rankingean duten posizioez gain, kosinuaren balioak berak ere erabil daitezke.

Esperimentuen emaitzetatik rankingak sortzeko prozesua Perl lengoaian programatu da.

VII.1.2.5 Emaitzak

Atal honetan egindako esperimentuen kopurua handia da, eta hemen emaitza interesgarrienak soilik azalduko ditugu. Emaitzen taulen egitura VII.1.1.1 taulan agerkidetza-esperimentuetarako emandako bera da.

Lehenik, VII.3 taulan bildu ditugu ohiko WSM ereduarekin egin ditugun esperimentuen emaitza onenak.

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
	t neurria	0,197	0,455	0,084	0,383
2000_MI_f3_%50	R_{BR}_N	0,185	0,416	0,217	0,259
	R_{BR}_V	0,242	0,431	0,108	**0,327**
	R_{BR}_NV	0,254	0,472	**0,233**	0,299
	R_{W_1}	0,262	0,472	0,197	0,310
	R_{W_2}	0,056	0,320	0,164	0,203
	Jac_N	(-0,036)	0,275	0,112	0,186
	Jac_V	(-0,018)	0,286	0,084	0,210
	Jac_NV	(-0,030)	0,275	0,113	0,186
	cos_N	(0,039)	0,306	0,139	0,198
	cos_V	(0,045)	0,316	0,083	0,237
	cos_NV	0,056	0,310	0,127	0,206
osoa_MI_%50	R_{BR}_N	0,185	0,401	0,213	0,251
	R_{BR}_V	0,223	0,391	0,088	0,307
	R_{BR}_NV	**0,270**	**0,473**	0,217	0,305
	R_{W_1}	0,227	0,400	0,097	0,307
	R_{W_2}	0,061	0,326	0,157	0,207
	Jac_N	(-0,054)	0,267	0,100	0,184
	Jac_V	0,117	0,345	0,072	0,277
	Jac_NV	(-0,015)	0,287	0,108	0,196
	cos_N	0,018)	0,294	0,133	0,193
	cos_V	(0,025)	0,304	0,074	0,232
	cosNV	(0,031)	0,300	0,121	0,201
osoa_f_f3_%50	JSdivN	(-0,057)	0,275	0,061	0,216
	JSdivV	(-0,007)	0,314	0,057	0,268
	JSdivNV	(-0,035)	0,296	0,055	0,247

VII.3 Taula: WSM ereduko antzekotasun distribuzionaleko neurrien emaitzak (emaitza nabarmenak izan dituzten esperimentuak soilik bistaratu ditugu). Esperimentuetako parametroak: a) WSM eredua eratzeko testuinguru-dokumentuen tamaina: 2 000 lerro (2000) edo testuinguru-lerro guztiak (osoa); b) bektoreen pisua: MI edo f; c) maiztasun-ataria bektorearen elementua izateko: f3 ($f \geqslant 3$) edo $\varnothing$ (ataririk ez); d) kontuan hartutako bektore-zatia: bektorearen erdia (% 50; beste esperimentu batzuetan, bektorearen lehen 2 000 edo 3 000 elementuak hartu dira kontuan, baina emaitzak okerragoak izan dira).

VII.1 taulako emaitzekin alderatuta, hauek dira nabarmentzeko alderdiak:

- R_{BR}_NV da, ia beti, emaitza onenak dituen neurria, Wulffek proposatutako R balioaren hedapenen, JS dibergentziaren eta, batez ere, arlo honetan hain aipatuak eta erabiliak diren Jaccard eta kosinuaren oso gainetik.

- τ_B-ren balioetan, lortzen da hobekuntza t neurriarekiko, baina, batik bat, esapide idiomatikoen $AP_{\texttt{id}}$-ren balioek agertzen dute hobekuntza nabarmena. Kualitatiboki, emaitza hori espero izatekoa da, teorikoki esapide idiomatikoak bailirateke idiosinkrasia semantiko nabarienekoak.

- Kolokazioen kasuan, emaitza onena aditzaren semantikarekiko konparazioa egiten duen R_{BR}_V neurriak lortu ditu (oro har, neurri guztietan gertatzen da _V neurketen $AP_{\texttt{col}}$-en balioa handiagoa izatea _N neurketena baino, eta alderantziz $AP_{\texttt{id}}$-ren balioetan); datu interesgarria da, koherentea delako kolokazioak erdikonposizionalak direlako ikuspegiarekin, baina, hala ere, ez du gainditzen t neurriaren emaitza.

- Esperimentuaren parametroak direla eta, bektoreen elementuak ordenatu eta iragazteko, MI da bektore-balioetarako neurri onena, eta emaitza onenak bektoreen elementu erdiak erabilita lortu dira, bektore osoak edo elementu-kopuru finko bat erabilita baino. Testuinguru-lerroak mugatuta zein testuinguru guztiak erabilita, antzeko emaitzak lortzen dira; esplikazio bat izan liteke bektore osoak ez erabiltzeak indargabetu egiten duela bigarren kasuan testuinguru-informazio hedatuagoa izatearen eragina.

Lemurrekin eta Infomapekin egindako neurketen emaitza onenak VII.4 taulan bildu ditugu. Indri eta KL indizeek izan dute portaera eraginkorrena; gainerakoek (Okapi, tf-idf, kosinua) beheragotik ibili dira, eta ez ditugu hona ekarri.

Hauek dira alderdi nagusiak:

- Lemurrekin egindako esperimentuen emaitzak Infomapekin egindakoen gainetik daude, alde nabarmenez.

- Emaitza onenak L2 modalitateko esperimentu batzuek dituzte, L1 esperimentuetakoak baino hobexeak baitira, oro har. Gogoan izan honelakoetan bigramaren testuinguru bakoitzeko kontsulta bat egiten dela,

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
	t neurria	0,197	0,455	0,084	0,383
L1	KL_rankN	0,135	0,404	0,263	0,236
	KL_rankV	0,290	0,497	0,136	0,371
	KL_rankNV	0,296	0,517	0,236	0,336
	Indri_rankN	0,161	0,422	0,240	0,259
	Indri_rankV	0,296	0,506	0,137	0,379
	Indri_rankNV	0,314	0,535	0,228	0,353
L2	KL_hit_erl_N	0,206	0,457	0,282	0,274
	KL_hit_erl_V	0,304	0,527	0,148	0,390
	KL_hit_erl_NV	0,308	0,551	**0,320**	0,332
	KL_rankN_hazt	0,204	0,455	0,286	0,272
	KL_rankV_hazt	0,308	0,539	0,132	0,415
	KL_rankNV_hazt	0,305	0,552	0,306	0,336
	Indri_hit_erl_N	0,213	0,463	0,265	0,285
	Indri_hit_erl_V	0,316	0,556	0,152	0,415
	Indri_hit_erl_NV	**0,322**	**0,566**	0,294	0,354
	Indri_rankN_hazt	0,214	0,456	0,261	0,280
	Indri_rankV_hazt	0,314	0,552	0,128	**0,431**
	Indri_rankNV_hazt	0,309	0,555	0,280	0,349
Infomap	rankN	(0,081)	0,293	0,045	0,264
	rankV	(0,257)	0,219	0,049	0,171
	rankNV	0,256	0,219	0,042	0,183
	cosN	(-0,029)	0,313	0,135	0,205
	cosV	(0,181)	0,406	0,089	0,324
	cosNV	0,169	0,424	0,174	0,284

VII.4 Taula: Antzekotasun distribuzionaleko IR erako esperimentuen emaitzak.

unigramen testuinguru-dokumentuen bildumaren kontra. Haien arteko aldeak ez dira oso handiak, eta ezin esanguratsutzat jo. Nolanahi ere, deigarria da hit-kopuru erlatiboan oinarritutako neurriak direla kasu gehienetan onenak (zeinetan ez baita kontuan hartzen bigramaren testuinguru bakun bakoitzari dagokion unigramaren dokumentuak

rankingean duen posizioa, lehen 500en artean agertzea baino ez da kontuan hartzen).

- AMen emaitzekin konparatuta, ebaluazio-metrika guztietan lortzen da hobekuntza nabaria, baina koska handiena esapide idiomatikoetan dugu. Horrek berresten eta areagotzen du VII.3 taulako WSM neurrietan behatu duguna: posible da esapide idiomatikoen erauzketa hobetzea, semantika karakterizatuta. Gainera, oraingoan badago neurri bat, Indri_rankV_hazt, VII.3 taulako R_{BR}_V neurriak bezala aditzaren semantika neurtzen duena, baina, horrek ez bezala, kolokazioekin t neurriak lortzen duen AP hobetzen duena. Emaitza hau oso interesgarria da, baieztatzen duelako kolokazioen erdikonposizionaltasuna, baita horren gakoa aditzaren semantika berezian dagoela ere; gainera, bat dator III.10 atalean aurkeztu ditugun Venkatapathy eta Joshiren (2005) emaitzekin. Aipagarria da neurri bera izenerako aplikatzen denean (L2_Indri_rankN_hazt), esapide idiomatikoekin askoz emaitza hobea duela aditzarenak baino. _N eta _V moduen arteko alde hori bertsua da gainerako neurrietan ere; beraz, esan liteke esapide idiomatikoen atzemateko, izenari begiratzea dela funtsezkoena.

- Oro har, Indri indizeak KL indizeak baino emaitza zertxobait hobeak ditu, $AP_{\texttt{id}}$-ren balioetan izan ezik, zeinetan KL_hit_erl_NV den neurri onena. Dena den, Indri eta KL neurketen arteko aldea ez da oso handia (lehen bi onenen artean, 0,02-0,03ko aldea dago, gehienaz ere), eta, KL kalkulatzea azkarragoa denez, aztertzekoa litzateke, aplikazio erreal bati begira, alde horrengatik Indri indizearen aldeko hautua egitea merezi duen.

VII.1.3 Malgutasun morfosintaktikoaren neurketa, erreferentzia-portaera batetiko distantziaren bidez

VII.1.3.1 Aldakuntza morfosintaktikoen hautaketa

Malgutasun morfosintaktikoaren neurketan kontuan hartzen diren aldakuntzen espektroak hizkuntzaren ezaugarrien araberakoa izan behar du nahitaez. Horrek berekin dakar euskararako diseinu-lan bat egin behar izatea, gure hizkuntzan malgutasunaren adierazle diren ezaugarriak, hau da, aldakuntzak, detektatu eta neurtzeko.

Aditz-lokuzioen aldagarritasuna aztertzean, Urizarrek dio (Urizar, 2012: 129-134) aldagarritasun-maila asko aldatzen dela esapide batzuetatik besteetara. Adibidez, lokuzioaren buru den aditzak, aditz jokatugabe gisa, flexio

guztiak har ditzake (*gogora ekarri/ekar/ekartzen/ekarriko...*), baina, aditz jokatu gisa, pertsona-murrizketak izaten ditu, lokuzio-motaren arabera. Esaterako, *min izan* lokuzioak singularreko 3. pertsona hartu behar du nahitaez objektutzat (*bihotzean min dut*, baina **min ditu/zintuen...*). *atsegin ukan* motako batzuek ez dute horrelako murrizketarik (*atsegin/maite/gorroto zaitugu/zintuztedan/banindute*), baina batzuek bai, ordea (**uste ditu*, **plazer zaituztegu*). Beste zenbait horrelako murriztapen-klase aipatzen ditu.

Bestetik, aditz-osagaiari laguntzen dion sintagmak flexio-aukera murritzagoa onartzen duela dio Urizarrek. *Min egin* eta *adarra jo* motetako aditz-lokuzio askotan, izenak partitiboa har dezake (*ez dizu minik egingo; ez adarrik jo!*), baina beste batzuetan ez (*alde egin, hots egin, men egin, zin egin*). Urizarrek Etxepareren (2003) iritzia dakar, zeinen arabera aditz-esapide telikoetako izenek partitiboa onartzen duten[7] (*eztul egin, zirkin egin, huts egin*), eta ez-telikoetakoek, berriz, ez, hala nola aspektu-ekintza adierazten dutenak (*bultza egin, laster egin*) eta lorpena adierazten dutenak (*bat egin, eztanda egin, topo egin*).

Beste aldagarritasun-mota bat aditz-lokuzioaren osagaien ordena eta jarraitutasunarekin dago erlazionatua. Urizarren arabera, aditz-lokuzioek aldagarritasun handia erakusten dute oro har, eta hein handi batean, dezenteko malgutasunez egokitzen bide dira mintzagai/galdegai mugimenduetara, perpaus-osagai askeak balira bezalatsu. Horren erakusgarritzat eman genitzake *lo egin* eta *lan egin* aditz-lokuzioen adibide hauek:

(17) **Egin lo** eta **hartu atseden**

(18) Bart gauean gaizki **egin** dut **lo**

(19) Non **egiten** duzu **lan**?

Bestetik, Urizarrek ohartarazten gaitu tradizioan zurrun erabili diren aditz-lokuzio batzuk (*alde egin, hots egin*) malgutasun handiagoa erakusten hasi direla gaur egun. Hain zuzen ere, ausart gintezke esatera hauek direla, edo izan direla, ohiko erabilerak[8]:

(20) *Mesedez, ez **egin alde** → Mesedez, ez **alde egin**

(21) *Nork **egin** dizu **hots**? → Nork **hots egin** dizu?

Hala ere, gaur egungo corpusetan ez da miraria horrelakoak aurkitzea:

[7] Telikotasuna aditzaren aspektu lexikoaren ezaugarri bat da, eta aditza ekintza-amaierarekin lotua izatea adierazten du.

[8] Esaterako, *Orotariko Euskal Hiztegi*ko adibideetan, *alde egin* edo *aldegin* ageri dira, baina ez da *egin alde* ordena duen adibiderik (*-en alde egin* lokuzioarenak alde batera utzita); orobat *hots egin*.

(22) Arineketan **egin** zuen **alde** (*Agirre zaharraren kartzelaldi berriak*, Koldo Izagirre)

(23) -Mirestekoa duk! - behin berriz **egin** zuen **hots** Agaramonteko kondeak (*Mailuaren odola*, Aingeru Epaltza)

Martínez (1996), Oyharçabal (2006) eta Odriozolaren (2010) ikuspegiak aurkezten ditu Urizarrek, agerian uzteko alderdi horiekiko malgutasuna aldakorra dela, aditz-lokuzioaren arabera ez ezik, euskalkiaren arabera ere bai. Gakoa bide da osagarriak aditzarekiko duen inkorporazio-maila edo, bestela esanda, zenbateraino den aditz-paradigmara sartzeko bidean; zenbat eta inkorporazio-maila handiagoa izan, hainbat eta aldakuntza gutxiago onartu.

Urizarrek aztertzen dituen aditz-lokuzioen aldagarritasun-kasuen artean ez dago aditzari laguntzen dion izenaren hedapenen ondoriozko malgutasuna. Horren arrazoia bide da Urizarrek horrelako hedapenak ez dituela aditz-lokuzioaren agerpentzat jotzen (Urizar, 2012: 105); horren arabera, *lan bat egin dut* eta *egin dudan lana* ez lirateke *lan egin* konbinazioaren aldakuntzak. II.4 atalean landu dugu arazo hau, eta argudiatu dugu horrelakoak aldakuntzatzat hartzea gure ikerkuntzan. Gainera, kontuan hartu behar da kolokazioen aldagarritasuna lantzen duen atalean (Urizar, 2012: 97), horrelako aldakuntzen adibideak aipatzen dituela Urizarrek (*izerdi hotza egin, bota dudan izerdia*).

Ikerkuntza honetan aditz-lokuzioetan ez ezik kolokazioetan ere interesatuta gaudenez, bistan da aldakuntza horiek ere aztergaitzat aitortu behar ditugula, zeren, esapide idiomatikoen eta kolokazioen a priorizko ezagutza ez dugunez, ezin baitugu bereizketarik egin izenaren hedapen bat aldakuntzatzat jotzeko ala ez. Beraz, modifikatzailea(k) gehitzea eta erlatibizazioa dira `izena+aditza` konbinazioen izenaren hedapen errelebanteak.

Esperimentuak diseinatzeko kontuan hartu dugun beste informazio-iturri bat EDBLko aditz-lokuzioen gauzatze-eskemak dira, lokuzioen deskribapena egiteko erabiltzen diren zehaztapenak (Urizar, 2012: 222-231). Horien bidez zehazten dira osagai bakoitzaren *flexio-murriztapenak*, euren aldagarritasun morfologikoaren berri emango dutenak, eta osagaien *hurrenkera-bariazioak*; azkenik, bada eremu berezi bat, *ziurtasuna*, osagaien agerkidetza aditz-lokuzioaren agerpenaren seinale ziurtzat jo daitekeen ala ez zehazten duena.

Esaterako, hauek dira *leher egin* eta *jaramon egin* lokuzioen gauzatze-eskemak:

```
12 ZIUR ([-],[%])
leher egin
```

```
12 ZIUR ((KAS=ABS ETA MUG=MG) EDO KAS=PAR,[%])
1+2 ANBI ((KAS=ABS ETA MUG=MG) EDO KAS=PAR,[%])
2*1 ANBI ((KAS=ABS ETA MUG=MG) EDO KAS=PAR,[%])
jaramon egin
```

Eskema horiek hau adierazten dute:

- *leher egin*
  - Ordena kanonikoan (12) bakarrik agertzen da, eta segurua da.
  - [-] ikurrak adierazten du izenak ez duela flexio-aldakuntzarik onartzen, hau da, erabateko murriztapena duela (beti *leher* forman behar du). [%] ikurrak, berriz, adierazten du aditzak edozein flexio har dezakeela.
- *jaramon egin*
  - Ordena kanonikoan (12), segurua da, izena ABS kasuan eta mugagabean bada (*jaramon*), edo partitiboan (*jaramonik*). [%] ikurrak adierazten du aditzak edozein flexio har dezakeela.
  - *jaramon* eta *egin* osagaien artean, elementu bat tarteka daiteke (1+2), baina orduan segida anbiguoa da, eta agerpena lokuzioa ez izatea gerta daiteke. Aurrekoan bezalako kasu- eta flexio-murriztapena dago.
  - Alderantzizko ordenan (*egin*+*jaramon*), elementu bat baino gehiago gerta daitezke osagaien artean (2*1); aurrekoan bezala, segida hori anbiguoa da. Aurrekoan bezalako kasu- eta flexio-murriztapena dago.

Bestetik, eremu jakin batek edozein balio onartzen duela adierazteko, izartxoa (*) erabiltzen da. Adibidez, hau da *maite izan* lokuzioaren lehen osagaiak (maite) duen flexio-murriztapena (Urizar, 2012: 227):

```
(KAS=ABS ETA MUG=MG) EDO (MAI=⋆ + (KAS=ABS ETA MUG=MG))
```

MAI=⋆ ezaugarriak adierazten du *maite* izenak edozein mailatako graduatzailea har dezakeela (*maiteago, maiteen, maiteegi*).

Ikerkuntza honetan, ez dugu helburutzat hartu UF baten agerpen idiomatikoak eta libreak (konposizionalak) bereiztea, eta, beraz, ziurtasuna ez da landu dezakegun alderdia.

Interesgarria da gauzatze-eskemen gainerako parametroen datu estatistikoak ematea: jarraitutasuna, ordena, eta izenaren zein aditzaren flexio-murriztapenak. EDBLko aditz-lokuzioen XML fitxategia prozesatuz, VII.5 taulako banaketak lortu ditugu jarraitutasuna eta ordenaren parametroetarako.

	Jarraitutasuna	**Ordena-aldaketa**
bai	114	693
ez	698	119
totala	812	812

VII.5 Taula: EDBLko aditz-lokuzioen *jarraitasuna* eta *ordena-aldaketa* parametroen araberako banaketa.

Datu horietatik, ondorio hauek atera ditugu:

- Aditz-lokuzio gehienak ez-jarraituak izan daitezke, hau da, osagaien artean elementu bat edo batzuk txerta daitezke. EDBLn ez da zehazten nolakoak izan daitezkeen txertatze horiek.
- Aditz-lokuzio gehienek ordena-aldaketa onartzen dute.

Bi parametro horiekiko, beraz, EDBLko aditz-lokuzio gehienak malguak dira. Kontuan hartzen badugu EDBLn jasotako gehienak esapide idiomatikoak edo euskarri-aditzen bidezko konbinazioak direla, eta gure ebaluaziolaginean, horiez gain, kolokazio asko ditugula (teorikoki behintzat horiek baino malguagoak direnak), datu horiek iradokitzen digute, nolabait, parametro horiek ez direla oso diskriminatzaileak izango UF-kategorien arteko karakterizaziorako. Nolanahi ere, esperimentalki egiaztatu beharreko irudipena da hori.

Bestetik, VII.6 taulan eman ditugu izenaren flexio-murriztapenaren ereduen banaketa.

Datu horiek erakusten diguten lehen egitate interesgarria da lokuzioen % 61ean izenaren forma aldaezina dela. Jarraitutasunaren eta ordena-aldaketaren parametroen aldean, badirudi flexio-murriztapenak idiomatikotasunarekin korrelazio handiagoa izateko zantzuak badaudela.

Izenaren flexio-murriztapenaren eredua	kopurua	%
[-]	499	61,00
(KAS=ABS ETA MUG=MG) EDO KAS=PAR	197	24,08
(KAS=ABS ETA NUM=S) EDO KAS=PAR	58	7,09
(KAS=ABS ETA MUG=MG) EDO (MAI=* + (KAS=ABS ETA MUG=MG))	14	1,71
(KAS=ABS ETA MUG=MG) EDO (MAI=KONP + (KAS=ABS ETA MUG=MG))	14	1,71
(KAS=ABS ETA MUG=MG) EDO (KAS=ABS ETA NUM=S) EDO KAS=PAR	9	1,10
KAS=ABS EDO KAS=PAR	5	0,61
[%]	4	0,49
(KAS=ABS ETA MUG=MG) EDO KAS=PAR EDO (MAI=KONP + (KAS=ABS ETA MUG=MG))	3	0,37
(KAS=ABS ETA NUM=P) EDO KAS=PAR	3	0,37
ADM=ADIZE + ((KAS=ABS ETA NUM=S) EDO KAS=PAR)	2	0,24
HAS_MAI [-]	1	0,12
(KAS=ABS ETA MUG=M) EDO KAS=PAR	1	0,12
(KAS=ABS ETA MUG=MG) EDO ((MAI=KONP EDO MAI=GEHI) + (KAS=ABS ETA MUG=MG))	1	0,12
(KAS=ABS ETA MUG=MG) EDO (KAS=ABS ETA NUM=P) EDO KAS=PAR	1	0,12
(KAS=ABS ETA NUM=P) EDO (KAS=GEN ETA NUM=P) EDO KAS=PAR	1	0,12
(NUM=S + KAS=INE) EDO (MUG=MG + KAS=INE)	1	0,12
AZP=SIN EDO KAT=ADT	1	0,12
NULL EDO ((KAS=ABS ETA NUM=S) EDO KAS=PAR)	1	0,12
KAS=ABS ETA MUG=MG	1	0,12
KAS=PAR	1	0,12
Totala	**818**	**100,00**

VII.6 Taula: EDBLko aditz-lokuzioen izenaren *flexio-murriztapena* parametroaren araberako banaketa; batura VII.5 taulakoa (812) baino handiago da (818), 6 lokuziotan 2 murriztapen-eredu zehaztu direlako.

Dena den, datua behar bezala interpretatzeko, aurretik honako ohar hau egin beharrean gaude.

Lehen begiratu batean, oso deigarria da kasuaren murriztapenean `ABS` eta `PAR` baino ez agertzea, behin baino ez baita bestelako kasu bat zehaztu: `(NUM=S + KAS=INE) EDO (MUG=MG + KAS=INE)` eskema, *bakean utzi* lokuzioari dagokiona, eta *bakean/baketan* aldakuntza deskribatzen duena. Pentsa genezake EDBLn hori dela izena `ABS` ez den kasuan agertzen den lokuzio bakarra, baina ez da horrela, hona hemen adibide batzuk: *argitara eman, aurpegira bota, begiz jo, burutik kendu, martxan jarri, kontuan hartu, kontutan hartu, telefonoz deitu*. Horien gauzatze-eskeman, izenerako, [-] agertzen da, hau da, izena ezin da erazagututako forman baino agertu. *Bakean utzi*ri bezalako tratamendua eman zekiokeen *kontuan hartu* lokuzioari, eta *kontutan hartu* aldakuntza flexio-murriztapenean deskribatu; horrelakorik ezean, bi lokuzioak dira sarrera EDBLn.

Aditzaren flexio-murriztapenak direla eta, VII.7 taulako datuek adierazten dute lokuzioen ehuneko handi bat malgua dela (% 74,85). Horiek horrela, ez dirudi aditzaren flexioa idiomatikotasuna karakterizatzeko parametro adierazgarria denik, eta, horrenbestez, erabaki dugu gure ikerkuntzan kontuan ez hartzea.

Gure hautua

Aurreko lerroetan bildutako informazioa kontuan izanik, jarraian azalduko ditugun aldakuntza-kasuak hautatu ditugu UF hautagaien malgutasun morfosintaktikoa karakterizatzeko.

`izena+aditza` konbinazioen `izena` osagaiaren aldakuntzetan jarri dugu arreta, izenaren ezker- zein eskuin-hedapen nagusiak eta mugatasun-aldakuntzak kontuan hartuz. Ezker-hedapenen artean, erlatibizazioa ere sartu dugu. Flexio-murriztapenen arloan, guretzat kasua ez da aldagai bat, konbinazioaren definizioaren beraren ezaugarri bat baizik. Beraz, mugatasun-aldakuntzak hartu ditugu kontuan; aparteko parametrotzat kontsideratu dugu mugatasuna, hedapenetatik bereiz; hartara, izen soilaren eta hedatuaren mugatasun-aldakuntzak parametro bakarrean konputatu dira. Azkenik, osagaien ordena ere parametrotzat hartu dugu.

Esaterako, hona hemen *liburua irakurri* konbinazioaren aldakuntza bakun batzuk:

- Izenaren ezker- eta eskuin-hedapenak:

Aditzaren flexio-murriztapenaren eredua	kopurua	%
[%]	628	74,85
AZP=SIN EDO (KAT=ADT ETA EZ(NORI))	52	6,20
AZP=SIN	31	3,69
AZP=SIN EDO (KAT=ADT ETA EZ(NORI) ETA NOR=HURA)	27	3,22
AZP=SIN EDO (KAT=ADT ETA EZ(NORI) ETA (NOR=HURA EDO NOR=HAIEK))	22	2,62
KAT=ADT	15	1,79
AZP=SIN EDO KAT=ADT	12	1,43
AZP=SIN EDO (KAT=ADT ETA EZ(NORI) ETA (NOR=HURA EDO NOR=HAIEK)) EDO (KAT=ADT ETA NORI=⋆)	11	1,31
AZP=SIN EDO (KAT=ADT ETA EZ(NORI) ETA NOR=HURA) EDO (KAT=ADT ETA NORI=⋆)	7	0,83
AZP=SIN EDO (KAT=ADT ETA (NOR=HURA EDO NOR=HAIEK))	4	0,48
AZP=SIN EDO (KAT=ADT ETA NOR=HURA)	4	0,48
AZP=SIN EDO (KAT=ADT ETA NORI=⋆)	4	0,48
KAT=ADT ETA EZ(NORI)	4	0,48
AZP=SIN EDO (KAT=ADT ETA EZ(NORI) ETA (NOR=HURA EDO NOR=HAIEK)) EDO (KAT=ADT ETA NORI=⋆)	2	0,24
AZP=SIN EDO (KAT=ADT ETA EZ(NORI)) EDO (KAT=ADT ETA NORI=⋆ ETA NOR=HURA)	2	0,24
KAT=ADI EDO (KAT=ADT ETA NOR=HURA)	2	0,24
Bestelakoak	12	1,43
Totala	**839**	**100,00**

VII.7 Taula: EDBLko aditz-lokuzioen aditzaren *flexio-murriztapena* parametroaren araberako banaketa; batura VII.5 taulakoa (812) baino handiagoa da (839), 23 lokuziok 2 murriztapen-eredu dituztelako, eta 2 lokuziok 3.

  - Determinatzailea: *liburu* ***bat*** *irakurri dut;* ***zenbat*** *liburu irakurri dituzu?*; *ez dut liburu* ***hori*** *irakurri*
  - Izenondoa: *liburu* ***interesgarria*** *irakurri nuen*
  - Izenlaguna: ***gustuko*** *liburuak irakurtzea*; ***italierazko*** *liburua irakurri*
  - Erlatiboa: ***irakurri dudan*** *liburua; anaiak* ***irakurritako*** *liburu batzuk*

  Esan gabe doa, hedapen bat baino gehiago konbina daitezke aldakuntza berean: *liburu interesgarri bat irakurri dut, lau liburu hauek irakurri ditut, anaiak irakurritako frantsesezko liburu eder batzuk.* Horrek eragina du aldakuntzen konputuan eta malgutasuna neurtzeko prozeduretan, hurrengo atalean azalduko dugunez.

- Mugatasuna: izenaren edo haren hedapenen mugatasun-aldakuntzak. Mugatasun-informazioa sintagmaren azken osagaiak darama. Kasu sinple batzuk:

  - *liburu****a****/liburu****ak****/*(zenbait) *liburu∅/liburu****ok*** *irakurri*
  - *egunkari****an****/egunkari****etan****/*(hiru) *egunkari****tan****/egunkari****otan*** *irakurri*
  - *liburu interesgarri****a*** */ interesgarri****ak*** *irakurri nuen/nituen*

  Argitzea komeni da letra lodiz nabarmendutako morfemak kasuaren eta mugatasunaren informazioa integratua dutela, baina ez dela hori parametro honetan kontuan hartzen dena, Eustagerren mugatasun-informazioa baizik, kasuaren informaziotik bereizia dena (`NUMS, NUMP, MG, PH`).

  Bestetik, mugatasun-informazioa, mugatzaileek ez ezik, bestelako determinatzaileek ere eramaten dute. Esaterako, determinatzailedun adibide hauetan parentesi artean eman dugu mugatasun-informazioa:

  - *liburu* ***bat*** *irakurri dut* (`MG`)
  - *ez dut liburu* ***hori*** *irakurri* (`NUMS`)
  - ***hiru*** *liburu* ***hauek*** *irakurriko ditut* (`NUMP`)

- Osagaien ordena (`IZE ADI / ADI IZE`): *liburua irakurri dut/irakurri dut liburua*

Aurrekoan bezala, izen-hedapena dagoenean ere ordena-aldakuntzak beha daitezke: *irakurri dut zure liburua, bai*; *aurten irakurri dut italierazko liburu hori*; *noiz irakurri duzu azken liburu interesgarria?* Dena den, bada hedapen bat ordena determinatzen duena: erlatibizazioa. Kasu horretan, sistemaren askatasun-graduen kopurua 1 da (`ADI IZE` egoera), eta egokiagoa dirudi horrelako ezker-hedapenak ez kontuan hartzea ordena-aldakuntzaren neurketan. Kontuan izan behar dugu aposizioan den perpaus erlatiboa (*bada liburu bat irakurri nahi nukeena*) ez dela izenaren hedapen sintagmatikoa; ez ditugu horrelakoak aztertu.

VII.1.3.2 Metodologia eta neurriak

Metodologiaren oinarria da aztergai dugun bigrama bakoitzaren portaera morfosintaktikoa erreferentzia-portaera batekin konparatzea. Horretarako, III.8 atalean azaldutako bi konparazio-bideak esperimentatu ditugu:

- Portaera orokorrarekiko konparazioa (Barkema, 1994a; Wulff, 2008; Fazly et al., 2009): kategoria-osaera bereko (gurean, `izena+aditza`) konbinazioen batez besteko portaerarekikoa.
- Osagaien portaerarekiko konparazioa (Bannard, 2007): konbinazioaren osagai batek beste osagaiaren kategoriako edozein hitzekin osatutako konbinazioen batez besteko portaerarekikoa.

Neurtzen den malgutasun-motaren arabera, bi faktore hauek hartu behar ditugu kontuan, eragin handia baitute neurketaren estrategian: a) malgutasunarekin erlazionatutako aldakuntzek paradigma itxi bat osatzen duten ala ez; eta b) malgutasunaren noranzkoa. Bi horiek xehatuko ditugu jarraian.

Aldakuntzen paradigma

Aztergaitzat hautatutako aldakuntza guztiak ez dira, paradigma osatzeari dagokionez, ezaugarri berekoak.

Mugatasun-aldakuntzek paradigma itxi bat osatzen dute. Izen-sintagmak mugatasun-ezaugarri bat eta bakarra du kasu bakoitzean: `[MG | NUMS | NUMP | PH | 0 ]`[9]. Bestetik, `izen-aditz` ordenak bi aukera baino ez ditu `[IZE ADI | ADI IZE]`), paradigma itxi bat osatzen dutenak.

[9]Teorian, `0` balioak ez luke paradigman agertu behar. Izan ere, balio hori mugatasun-informaziorik ez duen osagaiari dagokio, eta neurtu behar dugun mugatasun-aldakuntza mugatasun-informazioa duen osagaiarena da, sintagmaren eskuinen den osagaiarena, ale-

Horrrenbestez, mugatzailearen eta ordenaren kasuan, aldakuntza-moten probabilitateen batura 1 da, eta, ondorioz, mugatzailearekiko eta ordenarekiko malgutasunak neurketa banatan kalkula daitezke.

Aldiz, izenaren ezker- eta eskuin-hedapenek ez dute paradigma itxi bat osatzen, determinatzaileak edo bestelako modifikatzaileak sintagma berean konbina baitaitezke. Wullfek ere aipatzen du aldakuntza-mota honen ezaugarri hori (Wulff, 2008: 91-92). VII.4 irudian bildu ditugu *adostasuna lortu* bigramaren izenaren hedapen-moten aldakuntza konbinatu batzuk.

Horrek berekin dakar aldakuntza bakoitza independenteki neurtu behar izatea. Neurketa horietako bakoitzean, hedapenaren presentzia/absentziaren araberako banaketa hartzen dugu kontuan. Hedapen horiekiko malgutasunaren balio orokor bat nahi bada, neurketa horien emaitzak konbinatu egin behar dira; gure esperimentuetan, horien batezbestekoa kalkulatu dugu.

Malgutasunaren noranzkoa

Bestetik, esan dugu, malgutasuna neurtzeko, konbinazioaren portaeratik erreferentzia-portaera batera dagoen "distantzia" kalkulatzen dugula. Baina, distantziaz gain, bada "noranzkoa" ideia ere. `DET` aldakuntzen kasuan, adibidez, konbinazioaren izena `DET` batez modifikatzeko probabilitatea konparatu behar da probabilitate orokor batekin, eta bien arteko distantzia kalkulatu.

Argi dago ez dela gauza bera konbinazioaren `DET` probabilitatea probabilitate orokorra baino handiagoa edo txikiagoa izatea; lehen kasuan, konbinazioa malguagoa da, eta, bigarrenean, zurrunagoa. Distantziaren modulua bakarrik neurtuz gero, ez genituzke bi kasu horiek bereiziko, eta malgutasunaren neurketa okerra litzateke.

Horrek eragina du, Wullfek ohartarazi bezala, malgutasuna konputatzeko erabil daitezkeen neurrietan. Izan ere, neurri batzuek ez dute noranzkoa kontuan hartzen. Esaterako, III.8 atalean azaldu genituen neurrietatik:

- Fazly et al.-ek (2009) darabilten KL dibergentziak eta Wulffren (2008) SSD neurriak (*sum of squared deviations*) distantzia neurtzen dute, baina ez zein noranzkotan gertatzen den.
- Bannardren (2007) CPMIk (*conditional pointwise mutual information*) ondo jasotzen du noranzkoa, eta orobat Barkemaen (1994a) neurriak.

gia. Esaterako, *zenbait liburu eder irakurri* perpausean, *liburu* osagaiak ez du mugatasun-informaziorik (`0`), hori *eder* azken osagaiak daramalako (kasu honetan, `MG`), Baina, batzuetan, Eustagerren irteeran badira kasu batzuk non azken osagaiak ez duen mugatasun-informaziorik. Horregatik agertzen da balio hori paradigman.

aldakuntza	
	adostasun **handiena** lortu
	adostasun **sozial zabala** lortzea
	adostasun **bat** lortzea
	adostasun **minimo bat** lortu
nolabaiteko	adostasuna lortzea
erabateko	adostasuna lortzea
gutxieneko	adostasuna lortu
erakundeen arteko	adostasuna lortzea
independenteen	adostasuna lortuko
alderdien	adostasuna lortu
inguruko hainbat	adostasun lortu
oinarrizko	adostasun **zabalak** lortzea
horretarako	adostasun **zabala** lortzen
printzipioekiko	adostasun **zabala** lortzea
ustezko	adostasun **politikoa** lortzen
lortu gabeko oinarrizko	adostasun **maioritarioak**
batasunerako	adostasun **hori** lortuko
beste	adostasun **bat** lortu
gaiarekiko	adostasun **bat** lortzea
lortzen diren	adostasun **guztiak**
lortutako	adostasun **zabala**
lortutako	adostasun **politiko handiak**
lortutako	adostasun **hori**
lortutako gutxieneko	adostasuna

VII.4 Irudia: *adostasuna lortu* bigramaren aldakuntza batzuen adibideak (izenaren ezker- eta eskuin-hedapenak).

- Wulffek (2008) bere H_{rel} (entropia erlatiboa) neurriaren egokitzapen bat proposatu du, hain zuzen ere, noranzkoa kontuan hartzeko: *directional entropy* edo entropia direkzionala.

Gure esperimentuetan, kasu bakoitzean baliagarri diren aurreko neurriak probatu ditugu, zehaztapen hauekin:

- Mugatasunaren eta ordenaren kasuan, ez dugu malgutasunarekin erlazionatutako noranzkorik definitu. Arrazoia da ez dagoela erabat argi zer erlazio duen malgutasunak, esaterako, mugatasun-paradigma-

ko aukera jakin bakoitza errefererentzia-portaeran baino gehiago edo gutxiago agertzearekin. Printzipioz, pentsa liteke mugagabea maizago agertzea malgutasunik ezaren seinale izan litekeela; baina badaude bigrama batzuk non izenak maiz zenbatzaile zehaztugabea hartzen duen (`DET`+*lagun hil/gol sartu/euro balio*), eta ez dira horregatik idiomatikoagoak. Ordenaren kasuan ere, euskara ordena libre samarrekoa izaki, ez dago argi `ADI IZE` ordena erreferentzia-portaeran baino probableagoa izatea idiomatikotasunik ezaren agergarria den (gogoan izan erlatibodun agerpenak ez direla kontatzen ordena-aldakuntzetan).

- Barkemaren distantzia portzentuala erabili ordez, RFR (*relative frequency ratio*) erabili dugu, hau da, maiztasun erlatiboen arrazoia. Horrelako konparazioak egiteko, kendura baino neurri estandarragoa da.

- KL dibergentziaren kasuan, bigramaren portaera konparazio-portaeratik zein aldetatik urruntzen den gorde egiten dugu aldagai batean, eta, horren arabera, KL dibergentziaren balioari zeinu positiboa edo negatiboa ematen diogu ($-1, +1$ eskala bat eratuz). Balio negatiboek adierazten dute bigrama erreferentzia baino malgutasun txikiagokoa dela.

Behatu gabeko aldakuntzen tratamendua

Batzuetan, aldakuntza baten maiztasuna 0 da, eta behatu gabeko gertakarien arazoari aurre egin behar zaio. Izan ere, behatzen ez den aldakuntza baten probabilitatea zero denez, aldakuntza horri dagokion osagaia hutseratu egiten da neurriaren konputuan.

KL dibergentzia III.55 ekuazioan nola adierazi dugun kontuan izanik, bigramaren aldakuntza bat behatzen ez bada ($P(pt_k|v, n) = 0$), hitzarmena da $0 \log 0$ adierazpenaren balioa 0tzat hartzea[10]. Orobat gertatzen da Bannardek darabilen CPMIren III.59 ekuazioko $p(x|y, z)$ gaia 0 denean.

Arazoari ezikusia egitea da aukera bat. Den den, behatu gabeko aldakuntzen probabilitatea 0 izatea saihesteko, beste aukera bat da probabilitatea estimatzea, eta hori egiten denean, *smoothing* edo *leuntze* izeneko teknika aplikatzen dela esan ohi da.

CPMIren kasuan, Bannardek (2007) Laplace estimatzailea aipatzen du bere artikuluan. Metodo honek *Laplaceren segidaren araua* du oinarrian (Zabell, 1989), eta "bat gehitu" (*adding one*) izenaz ere ezaguna da (Manning eta Schütze, 1999: 202-203). Laplace-Bayes estimatzailea ere esaten zaio,

[10] http://en.wikipedia.org/wiki/Kullback-Leibler_divergence

Bayes estimatzailearen baliokidea baita, gertakariak a priori ekiprobableak direla joz gero[11].

Honela defini dezakegu hurrengo behaketan gertakari bat behatzeko probabilitatea[12]:

$$P(z=1) = \frac{n_1 + 1}{n_0 + \cdots + n_{K-1} + K} \ , \tag{VII.3}$$

non K gertakari kopurua den (gure kasuan, aldakuntza-kopurua), eta n_i, aldakuntza bakoitzaren maiztasuna. KL dibergentziarekin ere horren baliokide diren prozedurak proposatu dira[13].

Beraz, aldakuntza baten paradigmetako item baten maiztasuna 0 denean, banaketen leuntzea aplikatu dugu, Laplace-Bayes estimatzailea erabiliz.

VII.1.3.3 Aldakuntzen detekzioa corpusean

Bigramaren portaera ezagutzeko, bigrama bakoitza aldakuntza-mota bakoitzean agertzen den maiztasuna kontatzen dugu corpusean. Bestetik, malgutasuna kalkulatzeko erabili ditugun bi konparazio-bideen portaera ezagutzeko, bi prozedura hauek erabili ditugu:

- Portaera orokorra: aldakuntza-mota bakoitzaren datu orokorrak atera ditugu, edozein `izena`+`aditza` konbinaziotan zenbat aldiz agertzen den kontatuta eta `izena`+`aditza` konbinazioen agerpen guztiez zatituta.

- Osagaien portaera: gauza bera, baina edozein `izena`+`aditza` izan beharrean, bigramaren izena edozein aditzekin konbinatuta (eta alderantziz). Esaterako, *liburua irakurri* bigramaren kasuan, *liburua*+`ADI` eta `IZE`+*irakurri* konbinazioen batezbestekoak.

Hiru portaera horiek ezagutzeko,V.2.2 azpiatalean Eustaggerren irteera prozesatzeko deskribatu genuen sistema aberastu egin behar izan dugu, oraingoan, bigramen osagaien informazioa ez ezik, haien inguruko tokenen informazioa ere behar dugulako. Perl script baten bidez neurketa honetarako behar dugun informazioa formatu-mota honetara ekartzen dugu lehenik:

[11] http://en.wikipedia.org/wiki/Rule_of_succession

[12] http://web.engr.oregonstate.edu/~tgd/classes/534/slides/part6.pdf; http://www.uniroma2.it/didattica/WmIR/deposito/estimation_handout.pdf

[13] http://mathoverflow.net/questions/72668

```
berripaper berripaper IZE ARR 0 0
honetan hau DET ERKARR INE MG
ez ez PRT EGI 0 0
nuen *edun ADL B1 NRHU NK_NI
atzo atzo ADB ADOARR 0 0
horri hori DET ERKARR DAT NUMS
buruzko buruz ADB ALGARR GEL 0
aipamenik aipamen IZE ARR PAR MG
ikusi ikusi ADI SIN_PART BURU 0
```

Ondoren, PUNT_PUNT arteko elementuak lerro bakarrean antolatzen ditugu.

```
berripaper_berripaper_IZE_ARR_0_0 honetan_hau_DET_ERKARR_INE_MG
ez_ez_PRT_EGI_0_0 nuen_*edun_ADL_B1_NRHU_NKNI
atzo_atzo_ADB_ADOARR_0_0 horri_hori_DET_ERKARR_DAT_NUMS
buruzko_buruz_ADB_ALGARR_GEL_0 aipamenik_aipamen_IZE_ARR_PAR_MG
ikusi_ikusi_ADI_SINPART_BURU_0 PUNT_PUNT
```

Hurrengo urratsa da lerro bakoitza prozesatzea, aurreko hiru portaerak deskribatzeko behar dugun informazioa detektatzeko. Ikerkuntza honetan, sistema sinple bat garatu dugu maiztasun handieneko hedapen-egituren araberako sintagmak atzemateko; horretara, baliabide bakunekin, ahalik eta estaldura handiena lortzea da helburua.

Azaleko sintaxiko murriztapen-gramatika bat garatu dugu, adierazpen erregularrak darabiltzana, portaera bakoitza detektatzeko egokitu duguna. Datu horiek lortzeko scriptak Perl lengoaian idatzi ditugu.

Detekzio-prozesuaren urrats nagusiak hauek dira:

1 `izena+aditza` ordenako konbinazioen izenaren eskuin- eta ezker-hedapenak. Bigramen forma kanonikoaren definizioan kasua sartu dugunez, aldakuntzak detektatzean kasuaren presentzia kontrolatu egin behar da, eta izenak edo haren eskuin-hedapenak bigramaren forma kanonikoan den kasu berean agertu behar du, aldakuntza bigramari esleituko bazaio.

   Erauzten diren osagaien etiketei `aur_` edo `ond_` ezartzen zaie aurretik. Eskuin-hedapenen kasuan, osagai flexionatuaren mugatasuna ere atzematen dugu (`mugat_`); ezker-hedapenak daudenean, egiaztatzen da aurrez eskuin-hedapenik atzeman den ala ez; ezezkoan, izenaren mugatasuna atzematen da. Ordena ere erregistratzen da (`ord_`). Izenaren

ezker-hedapena erlatiboa denean, urrats honetan ez ezker- ez eskuin-hedapenik erauzten da, ezta dagokion mugatasun-informazioa ere (horiek 3. urratsean erauzten dira).

2 `aditza+izena` ordenako konbinazioen izenaren eskuin- eta ezker-hedapenak. Erauzten diren osagaien etiketei `aur_` edo `ond_` ezartzen zaie aurretik. Aurrekoan bezalako irizpideak erabiltzen dira kasua, mugatasuna, erlatiboa eta ordena tratatzeko.

3 Izenaren erlatibodun ezker-hedapenak. Aditz trinkoen eta laguntzaileen gainekoak (`erlt_JOK`) zein jokatugabeen gainekoak (*-tako/-dako*; *-(r)iko*: `erlt_PART`). Lehen puntuan aurreratu bezala, izenaren ezkerzein eskuin-hedapenak ere erauzten dira urrats honetan. Hori egiteko arrazoiak zerikusia zuzena dute erlatiboaren ezaugarri bereziekin, laster azalduko dugunez.

4 Aldakuntza morfosintaktikorik gabeko bigrama-agerpen hutsak. Bigrama hutsen maiztasunak beharrezkoak dira aldakuntzen maiztasunekiko konparazioan erabiltzeko. Mugatasuna eta ordena ere erregistratzen da.

Erlatiboaren kasu berezia

Erlatiboaren bidezko ezker-hedapenen detekzioak ezaugarri bereziak ditu.

- Lehen adierazi dugu erauzten den sintagmak bigramaren forma kanonikoaren kasu berean agertu behar duela, bigramaren aldakuntzatzat jotzeko. Bada, ordea, aldakuntza bat non hori ezin den eskatu: erlatiboa. Izan ere, erlatibo-perpausaren buru den izenaren kasua "ezabatuta" dago. Esaterako, *atzo irakurri nuen liburutik bi ondorio atera ditut* esaldia, berez, *liburua irakurri* bigramaren aldakuntza da, baina *liburu* lema `ABL` kasuan dago. Horiek kontuan hartu nahi izanez gero, kasua ezin da baldintzatzat jarri. Gure esperimentuetan, horrela jokatu dugu, baina horrek albo-ondorio batzuk baditu: lema-konbinazio bat kasu desberdinez bi bigramatan badago (*egunkaria irakurri, egunkarian irakurri*), erlatiboaren kontaketak biei esleitzen zaizkie; ondorioz, biek `ERLT` kontaketa berberak dituzte, eta, seguru aski, bati bakarrik dagozkio (kasu honetan, `ABS` kasua duenari).

  Aurrekoak inplikazio handiagoak ditu bigramaren osagaien erlatiboarekiko portaera neurtu nahi denean, edo `izena+aditza` portaera oro-

korra zehaztean. Esaterako, *adarra jo* eta *adarretatik heldu* bigramen malgutasuna neurtzeko erreferentziatzat `adar_ABS+ aditza` eta `adar_ABL+ aditza` konbinazioen portaera hartzen badugu, ikusiko dugu, aurreko *egunkaria irakurri, egunkarian irakurri* bigramekin gertatzen denaren antzera, `ERLT` kontaketa bat bera dela bientzat, ez baitugu `ABS` eta `ABL` kasua baldintzatzat hartzerik `ERLT` aldakuntzaren agerpena esleitzeko. Portaera orokorraren aldakuntzen kontaketa egitean, kasua hartzen da gakotzat, baina `ERLT` aldakuntzaren kasuan, irtenbide bakarra da kasua ez kontsideratzea.

- Detekzio-urratsak azaldu ditugunean, adierazi dugu erlatiboaren bidezko ezker-hedapena duen izen baten bestelako ezker- zein eskuinhedapenak erlatiboa detektatzen den urrats berean erauzten ditugula. Esaterako, *irakurritako zenbait liburutan* eta *irakurri ditudan liburu ederrengatik* sintagmetatik `DET` eta `ADJ` aldakuntzak erauzi behar genituzke. Bada, hori horrela egiteko arrazoia hobeto uler dezakegu aurreko paragrafoan azaldutakoa kontuan hartuta. `ERLT` analisi bat atzematen denean, ez dugu izenaren kasua egiaztatu behar, ezkutatuta baitago. Beraz, izenaren ezkerrean edo eskuinean izan daitezkeen `DET`, `IZL` eta `ADJ` analisiak bakarrik interesatzen zaizkigu, eta azkenaren mugatasun-informazioa baino ez. Honelakoetan, erauzten diren izenaren ezker-hedapenak `aur_erlt_` ezaugarria daramate, eta `ond_erlt_` eskuinekoek.

- Erlatiboak anbiguotasun-arazoak ditu. Oro har, *-(e)n* atzizkia duten adizkiek (laguntzaileek zein trinkoek) `ERLT` (erlatiboa), `ZHG` (zehar-galdera) eta `MOS` (mendeko osagaia) etiketak izan ditzakete (`ERL MEN` etiketen ondoren). Gainera, azpikategoria hauetako adizkietan, adizki hutsaren analisia ere desanbiguazio-aukeretako bat da: `A3` (*dadin*), `B1` (*zen*), `B3` (*zatekeen*), `B5` (*zedin*) eta `B8` (*zitekeen*).

  Gure hipotesia da desanbiguazio ez-zuzenak, isiltasuna eta zarata sortzen dutenak, uniformeki banatuta daudela aditzen artean, hau da, aditz guztiek errore-tasa bera dutela. Hartara, kontsideratu dugu desanbiguazio automatikoak egindako lana ontzat hartzea, `ERLT` analisia duten hedapenak soilik zenbatzea, eta `ZHG` edo `MOS` analisiak erlatiboen kontaketetarako kontuan ez hartzea.

  Horiek horrela, ez da esan beharrik garrantzi handia duela aditz jokatuen gaineko erlatiboa eta jokatugabeen gainekoa bereiztea, a priori

bederen, bigarrena askoz seguruagoa baita. Horrexegatik bereizi ditugu `erlt_JOK` eta `erlt_PART` hedapenak.

VII.1.3.4 Aldakuntzen kontaketa eta neurrien kalkulua

Aurreko atalean deskribatutako detekzio-prozesuan erauzten dugun informazioaren adibide gisa, VII.5 irudian dago lehen bistaratutako *adostasuna lortu* bigramaren aldakuntzetatik erauzten ditugun hedapenen informazioa.

aldakuntza		hedapen-motak
	adostasun **handiena** lortu	ond_ADJ
	adostasun **sozial zabala** lortzea	ond_ADJ_ADJ
	adostasun **bat** lortzea	ond_DET
	adostasun **minimo bat** lortu	ond_ADJ_DET
nolabaiteko	adostasuna lortzea	aur_ADB_GELN
erabateko	adostasuna lortzea	aur_ADJ
gutxieneko	adostasuna lortu	aur_DET_GELN
erakundeen arteko	adostasuna lortzea	aur_IZE_GELN_ADJ
independenteen	adostasuna lortuko	aur_ADJ_GELN
alderdien	adostasuna lortu	aur_IZE_GELN
inguruko hainbat	adostasun lortu	aur_IZE_GELN_DET
oinarrizko	adostasun **zabalak** lortzea	aur_ADJ / ond_ADJ
horretarako	adostasun **zabala** lortzen	aur_DET_GELN / ond_ADJ
printzipioekiko	adostasun **zabala** lortzea	aur_IZE_GELN / ond_ADJ
ustezko	adostasun **politikoa** lortzen	aur_ADJ / ond_ADJ
lortu gabeko oinarrizko	adostasun **maioritarioak**	aur_ADB_GELN_ADJ / ond_ADJ
batasunerako	adostasun **hori** lortuko	aur_IZE_GELN / ond_DET
beste	adostasun **bat** lortu	aur_DET / ond_DET
gaiarekiko	adostasun **bat** lortzea	aur_IZE_GELN / ond_DET
lortzen diren	adostasun **guztiekin**	erlt_JOK / ond_erlt_DET
lortutako	adostasun **zabala**	erlt_PART / ond_erlt_ADJ
lortutako	adostasun **politiko handiak**	erlt_PART / ond_erlt_ADJ_ADJ
lortutako	adostasun **hori**	erlt_PART / ond_erlt_DET
lortutako gutxieneko	adostasuna	erlt_PART / aur_erlt_ADJ

VII.5 Irudia: *adostasuna lortu* bigramaren aldakuntza batzuk (izenaren ezker- eta eskuin-hedapenak), eta erauzten diren hedapenak.

Hedapen-moten kontaketa eginez,VII.8 taulako emaitzak lortzen ditugu.

Mugatasuna eta ordena direla eta, VII.9 taulan bistaratu dugu *adostasuna lortu* bigramaren aldakuntza-mota horien kontaketa.

Antzeko erauzte-prozesua inplementatu dugu erreferentzia-portaeretarako. Esaterako, VII.10 taulan daude *adostasuna*+`aditza` konbinazioen hedapenen kontaketak. Izenaren kasua kontuan hartzen da aldakuntza-moten

adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	aur_ADB_GELN	15
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	aur_ADB_GELN_ADJ	1
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	aur_ADJ	56
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	aur_ADJ_GELN	18
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	aur_DET	9
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	aur_DET_GELN	51
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	aur_IZE_GELN	112
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	aur_erlt_ADJ	2
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	erlt_JOK	9
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	erlt_PART	90
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	hutsa	559
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	ond_ADJ	176
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	ond_ADJ_ADJ	6
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	ond_ADJ_DET	10
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	ond_DET	49
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	ond_erlt_ADJ	7
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	ond_erlt_ADJ_ADJ	1
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	ond_erlt_DET	6

VII.8 Taula: *adostasuna lortu* bigramaren hedapen-moten kontaketa.

kontaketan; hortaz, *adostasunera heldu* bigramaren kasuan, konparazioa egiten da `adostasun_ABL_aditza` gakoa duten konbinazioekin.

Konparazioa `izena+aditza` konbinazioen portaera orokorrarekiko egiten denean:

- Aldakuntzak izenaren kasu bakoitzerako erauzten dira (birdeklinazioa kontuan izanik).
- Erlatibo bidezko ezker-hedapenetan, ez dugu aukerarik hedapena izenaren kasu bati lotzeko. Beraz, `NULL` kasuari esleitu zaizkio `aur_erlt_` eta `ond_erlt_` hedapenak.

Lehen esan bezala, hedapenen kasuan hedapen bakunekiko malgutasuna bereiz neurtu behar dugu. Hedapen bakun hauek neurtzea erabaki dugu:

- `DET`: aurrekoa zein ondorengoa (`DET` duten `aur_` hedapenak eta `DET` amaiera duten `ond_` hedapenak; `_erlt_` motako hedapenak barne).

  `DET` duten `aur_` hedapenen kasuan, `_DET` amaiera behar da `DET` hedapentzat hartua izateko. Esaterako, VII.5 irudiko *gutxieneko adostasuna*

adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	mugat_MG	42
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	mugat_NUMP	47
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	mugat_NUMS	395
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	mugat_erlt_0	11
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	mugat_erlt_MG	3
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	mugat_erlt_NUMP	19
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	mugat_erlt_NUMS	66
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	mugat_hutsa_MG	4
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	mugat_hutsa_NUMP	64
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	mugat_hutsa_NUMS	491
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	ADIIZE	38
adostasuna_adostasun_IZE_ARR_ABS_NUMS<>lortu_ADI	IZEADI	1005

VII.9 Taula: *adostasuna lortu* bigramaren mugatasun- erta ordena-aldakuntzen kontaketa.

adostasun	ABS	ADITZA	aur_ADB_GELN	31
adostasun	ABS	ADITZA	aur_ADB_GELN_ADJ	1
adostasun	ABS	ADITZA	aur_ADJ	129
adostasun	ABS	ADITZA	aur_ADJ_GELN	41
adostasun	ABS	ADITZA	aur_DET	22
adostasun	ABS	ADITZA	aur_DET_GELN	110
adostasun	ABS	ADITZA	aur_IOR_GELN	11
adostasun	ABS	ADITZA	aur_IZE_GELN	289
adostasun	ABS	ADITZA	aur_IZE_GELN_ADJ	2
adostasun	ABS	ADITZA	aur_erlt_ADJ	13
adostasun	ABS	ADITZA	aur_erlt_DET	2
adostasun	ABS	ADITZA	erlt_JOK	68
adostasun	ABS	ADITZA	erlt_PART	133
adostasun	ABS	ADITZA	hutsa	1254
adostasun	ABS	ADITZA	ond_ADJ	390
adostasun	ABS	ADITZA	ond_ADJ_ADJ	11
adostasun	ABS	ADITZA	ond_ADJ_DET	23
adostasun	ABS	ADITZA	ond_DET	107
adostasun	ABS	ADITZA	ond_erlt_ADJ	24
adostasun	ABS	ADITZA	ond_erlt_ADJ_ADJ	2
adostasun	ABS	ADITZA	ond_erlt_DET	12

VII.10 Taula: *adostasuna*+`aditza` konbinazioen hedapen-moten kontaketa.

eta *horretarako adostasuna* aldakuntzak `DET` kategoriako lema banaren gaineko `IZL` hedapenak dira, eta ez *adostasun*en `DET` motako ezker-hedapenak. Aldiz, *inguruko hainbat adostasun lortu* aldakuntzatik, *hainbat*i dagokion `DET` hedapen bakuna kontatzen dugu (bi hedapen ditugu: `IZL` eta `DET`).

- `ADJ`: ondorengoa (hau da, `ond_ADJ` duten hedapenak); hedapen bereko `ond_ADJ_ADJ` hedapenak `ADJ` hedapen bakartzat jotzen ditugu, `ADJ` bat edo bi hartzea ez baitugu konbinazioa malguagoa izatearekin lotzen.

- `IZL`: izenlagun-kategoriako ezker-hedapena (hau da, `aur_ADJ`) zein `GEL` edo `GEN` bidez (`GELN`) osatutako izenlagunak, oinarrian, `IZE`, `ADJ`, `IOR`, `DET` edo `ADB` izan dezaketenak. Ikerketa honetan, ez dugu esanguratsutzat jo `GELN` motako ezker-hedapenen barne-egitura. Esaterako, *Europako Batzordeko Liburu Berdean* eta *Maiatzeko Plazako amen mugimenduari buruzko liburuaren frantsesezko itzulpena* sintagmetan, *liburu* izenaren `aur_GELN` egiturako bi hedapen zenbatzen ditugu, horien barne-egitura malgutasuna kuantifikatzeko ezaugarri esanguratsua ez dela kontsideratuz.

- erlatiboa: `erlt_JOK` (aditz trinkoaren gainekoa) eta `erlt_PART` (partizipioaren gainekoa). Erlatiboarekiko malgutasuna bitara neurtu dugu: lehenik, bi hedapenak batera kontuan hartuz (`erlt_JOK+erlt_PART`); eta, bestetik, jokatugabeen edo partizipioaren gainekoa soilik kontatuz.

Hedapen bakun bakoitzarekiko malgutasuna neurtzeko, bi probabilitate-banaketa konparatu behar ditugu: bigramarena eta erreferentzia-portaerako konbinazio ereduarena. Esaterako, hedapen bakun batekiko malgutasuna neurtzeko, hedapen bakun hori agertzen den hedapen-moten maiztasunak batzen dira (erauzte-prozesuak bermatzen du agerpen disjuntuak direla), eta hedapen hori ez dutenen maiztasunak ere bai (`DET`-en kasuan, hedapenik gabeko `hutsa` balioa dutenak eta `DET` ez duten gainerako hedapenak).

Horretara, *adostasuna lortu* bigramaren portaera eta *adostasuna*+`aditza` konbinazioen portaera `DET` hedapen bakunarekiko konparatzeko, VII.8 eta VII.10 tauletako informazioa prozesatzen dugu, eta VII.11 taulan dauden banaketak kalkulatzen.

Horrelako informazioa erabiliz, bigramaren malgutasuna neur dezakegu. Esaterako, VII.12 taulan *hanka sartu* esapide idiomatikoaren, *adostasuna lortu* kolokazioaren eta *liburua argitaratu* konbinazio librearen `DET` hedapenarekiko malgutasunaren neurriak eman ditugu, portaera-erreferentziatzat

	DET	ez_DET
adostasuna lortu	74	969
adostasuna+aditza	164	2 220

VII.11 Taula: *adostasuna lortu* bigramaren eta *adostasuna*+aditza konbinazioen DET hepapenarekiko banaketak.

osagaien portaera erabiliz. "_izena" daramaten neurriak kalkulatzeko erreferentzia-portaeratzat, *adostasuna*+aditza moduko konbinazioen batez besteko portaera erabili dugu, hau da, *bigramako izena*+edozein aditz konbinazioen portaera; esaterako, *hanka hautsi, hanka busti, hanka lotu*; *adostasuna bilatu, adostasuna landu, adostasuna adierazi*; *liburua irakurri, liburua galdu, liburua zaindu.* "_aditza" daramatenen kasuan, berriz, edozein izen+*bigramako aditza* konbinazioen portaerarekin egin da konparazioa; adibidez, *gola sartu, aldaketak sartu, muturra sartu*; *helburuak lortu, baimena lortu, askatasuna lortu*; *aldizkaria argitaratu, diskoa argitaratu, emaitzak argitaratu.*

	hanka sartu	adostasuna lortu	liburua argitaratu
RFR_DET_big_izena	0,12534	1,02256	1,32203
RFR_DET_big_aditza	0,04399	0,37068	1,48816
KL_div_DET_big_izena	−0,07189	0,00002	0,01803
KL_div_DET_big_aditza	−0,31053	−0,05311	0,03421
CPMI_DET_big_izena	−2,84942	0,02996	0,22049
CPMI_DET_big_aditza	−3,99001	−1,24995	0,33310

VII.12 Taula: *hanka sartu*, *adostasuna lortu* eta *liburua argitaratu* bigramen DET hedapenarekiko malgutasunaren neurri batzuk, portaera-erreferentziatzat osagaien portaera erabiliz.

Adibide horiek baliatuko ditugu erabilitako neurrien ezaugarriak azaltzeko, emaitzak egoki interpretatzearren.

- RFR (*relative frequency ratio*): bigramaren portaera erreferentziarekiko desbideratzen ez denean, balioa 1 da. RFR zenbat eta txikiagoa izan, hainbat eta malgutasun txikiagoa du (RFR $>$ 1 denean, erreferentzia baino malguagoa da).
- KL dibergentzia (KL_div) eta CPMI: balioa 0 denean, bi banaketen

artean ez dago alderik (konbinazioaren malgutasuna ez da erreferentziatik bereizten). `DET` parametroa direkzionala denez, kasu honetan balio negatiboek adierazten dute bigrama erreferentzia baino malgutasun txikiagokoa dela.

Neurrien emaitzek erakusten dute *hanka sartu* bigrama dela zurruntasun handiena duena izenak determinatzailea hartzeari dagokionez, *adostasuna lortu* dela eskala horretan hurrena, eta, azkenik, *liburua argitaratu* dela `DET` osagaia errazen hartzen duena, erreferentziatzat hartutako portaera bera baino malguagoa izateraino. Esan genezake hiru adibide horietan badela malgutasun morfosintaktikoaren eta idiomatikotasunaren arteko nolabaiteko korrelazio bat.

Hiru kasuetan, zurruntasuna nabarmenagoa da bigramaren portaera konparatzen dugunean dagokion aditzak beste izenekin osatzen dituen konbinazioen portaerekin (hau da, *hanka sartu* konbinazioaren portaera `izena`+*sartu* konbinazioen portaerarekin konparatuta).

Beste adibide bat jarriko dugu, mugatasun-aldakuntzekiko malgutasunaren kalkulua ilustratzeko, erreferentziatzat portaera orokorra erabiliz. VII.13 taulan, `izen_ABS`+`aditza` konbinazioen mugatasun-aldakuntzen konputua eman dugu.

ABS	mugat_MG	837 721
ABS	mugat_NUMP	544 750
ABS	mugat_NUMS	1 064 323
ABS	mugat_PH	1 960
ABS	mugat_hutsa_MG	1 568 315
ABS	mugat_hutsa_NUMP	654 027
ABS	mugat_hutsa_NUMS	1 334 492
ABS	mugat_hutsa_PH	6 338

VII.13 Taula: `izen_ABS`+`aditza` konbinazioen mugatasun-aldakuntzen konputua.

.

Nabaria den lehen alderdia da portaera orokorra definitzeko ez ditugula `mugat_erlt` motako aldakuntzak sartu. Horren arrazoia azaldua dugu dagoeneko: erlatibodun izen-sintagmen kasuan, ez du zentzurik kasuaren informazioa bigramaren kasuarekin bat datorren ala ez egiaztatzeak. Beraz, horrelakoak, mugatasun-aldakuntzetan ere, `NULL` kasuarekin erlazionatu ditugu. Baina, `ERLT` aldakuntzetan ez bezala, erabaki dugu gainerako `mugat`

aldakuntzekin ez konbinatzea, mugatasunaren informazioa bereizirik atxikitzeko. Egia da hori, berez, erlatiboarekin berarekin gertatzen dela, ez horren mugatasunarekin bakarrik, baina erlatiboaren kasuan ez dugu beste informaziorik, eta mugatasunaren kasuan, berriz, badugu `mugat_erlat` ez diren agerpenena.

VII.14 taulan ditugu VII.12 taulako hiru bigramen mugatasun-aldakuntzekiko malgutasunaren emaitzak, erreferentziatzat portaera orokorra erabiliz ateratakoak.

	hanka sartu	adostasuna lortu	liburua argitaratu
MSFlex_SSD_mugat_big_orok	0,28702	0,17931	0,08818
MSFlex_KL_div_mugat_big_orok	0,70742	0,51300	0,25577
MSFlex_CPMI_mugat_big_orok	19,51428	19,15057	25,19797
MSFlex_Hrel_mugat_big	0,38967	0,52916	0,67754

VII.14 Taula: *hanka sartu*, *adostasuna lortu* eta *liburua argitaratu* bigramen mugatasun-aldakuntzekiko malgutasunaren neurri batzuk, portaera-erreferentziatzat `izena+aditza` konbinazioen batez besteko portaera erabiliz.

Gogorarazi beharra dago mugatasun-aldakuntzen kasuan ez dugula noranzkoa kontuan hartu. Beraz, SSD eta KL dibergentziaren kasuan, balio handiek adierazten dute bigramaren mugatasunarekiko portaera urrun dagoela batez besteko portaera orokorretik; 0 balioa dutenean, ez dago desbideratze edo dibergentziarik bi banaketen artean. SSD eta KL neurriek iradokitzen dute badagoela korrelazio bat bigramaren idiomatikotasunaren eta mugatasun-malgutasunaren artean.

Aldiz, CPMI eta Wulffen H_{rel} neurriak alderantziz interpretatu behar dira (balio txikiak idiomatikotasunaren adierazgarri dira). CPMIk ez du islatzen *hanka sartu* eta *adostasuna lortu* bigramen arteko malgutasun-alderik dagoenik. Bestetik, Wulffen H_{rel} neurriak adierazten digu zenbateraino urruntzen den bigramaren mugatasun-portaera entropia maximoko banaketa batetik (non egoera guztiak ekiprobableak diren). Kasu honetan, balioak handituz doaz idiomatikotasuna gutxituz doan heinean.

Aldaeren detekzioa eta kontaketa bezala, neurrien kalkulua Perl scripten bidez egin dugu.

VII.1.3.5 Emaitzak

VII.15 taulan, izenaren `DET`, `ADJ` eta `IZL` hedapen-aldakuntzekiko malgutasun morfosintaktikoaren neurketen emaitzak ageri dira. Horiek aztertuta,

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
	t neurria	0,197	0,455	0,084	0,383
DET	RFR_orok	0,068	0,354	0,148	0,234
	KL_div_orok	(-0,036)	0,302	0,133	0,200
	CPMI_orok	0,074	0,369	**0,162**	0,240
	Hrel	0,088	**0,376**	0,155	0,248
	RFR_izena	(-0,043)	0,280	0,093	0,200
	RFR_aditza	**0,088**	0,371	0,135	**0,255**
	RFR_osag	(0,030)	0,325	0,120	0,224
	KL_div_izena	(-0,084)	0,251	0,070	0,185
	KL_div_aditza	0,070	0,338	0,122	0,234
	KL_div_osag	(-0,008)	0,286	0,100	0,202
	CPMI_izena	(-0,022)	0,289	0,099	0,204
	CPMI_aditza	0,072	0,361	0,128	0,248
	CPMI_osag	0,048	0,335	0,125	0,228
ADJ	RFR_orok	(0,032)	0,361	0,188	0,221
	KL_div_orok	(-0,003)	0,289	0,136	0,187
	CPMI_orok	(0,045)	**0,387**	**0,202**	0,237
	Hrel	(-0,054)	0,318	0,044	**0,298**
	RFR_izena	(-0,059)	0,286	0,085	0,210
	RFR_aditza	0,056	0,377	0,152	0,251
	RFR_osag	(-0,014)	0,312	0,116	0,217
	KL_div_izena	(-0,090)	0,254	0,065	0,192
	KL_div_aditza	(0,027)	0,324	0,104	0,229
	KL_div_osag	(-0,039)	0,276	0,080	0,200
	CPMI_izena	(-0,044)	0,296	0,091	0,216
	CPMI_aditza	**0,059**	0,378	0,157	0,250
	CPMI_osag	(0,025)	0,342	0,122	0,236
IZL	RFR_orok	0,108	0,387	0,103	0,291
	KL_div_orok	0,099	0,362	0,112	0,266
	CPMI_orok	0,097	0,375	0,102	0,278
	Hrel	0,115	0,386	0,099	**0,292**
	RFR_izena	**0,139**	0,390	0,145	0,269
	RFR_aditza	0,083	0,369	0,085	0,289
	RFR_osag	0,117	0,376	0,117	0,271
	KL_div_izena	0,137	**0,397**	**0,170**	0,272
	KL_div_aditza	0,066	0,334	0,074	0,263
	KL_div_osag	0,107	0,374	0,135	0,268
	CPMI_izena	0,122	0,376	0,143	0,254
	CPMI_aditza	0,074	0,366	0,091	0,281
	CPMI_osag	0,110	0,377	0,116	0,271

VII.15 Taula: Izenaren DET, ADJ eta IZL hedapen-aldakuntzekiko malgutasun morfosintaktikoaren neurketaren emaitzak.

ohar hauek egingo ditugu:

- τ_B-ren balioak urrun daude t neurriaren emaitzetatik; horrek adierazten du izenaren hedapenekiko malgutasunak ez duela korrelazio handirik agertzen idiomatikotasun-mailarekin. Areago, $AP_{\texttt{UF}}$-ren balioek erakusten dute, UFen erauzketa hutsean ere, agerkidetza-neurriaren azpitik dabiltzala malgutasun-mota honen neurriak.

- Hala ere, $AP_{\texttt{id}}$-ren balioetan hobekuntza nabaria lortu da t neurriarekiko. Hobekuntza hori nabarmenagoa da `ADJ` hedapenaren kasuan, eta `DET` zein `IZL` maila bertsuan dabiltza. Kolokazioen kasuan, berriz, t neurriaren azpitik dabiltza emaitza guztiak. Horren interpretazioa da esapide idiomatikoetan bakarrik dela esanguratsua izenaren hedapenarekiko zurruntasuna. Beraz, bi kategorien arteko bereizketan zeresana izan dezake malgutasun-mota honek.

- Erabilitako neurrien arteko aldeak direla eta, ez da besteei nabarmen gailentzen zaien neurririk. Dena den, emaitza esanguratsuak esapide idiomatikoei dagozkienez, CPMI neurria da horretan eraginkorrena (`IZL` hedapenaren kasuan izan ezik, non KL dibergentzia nagusitu den).

- Portaera orokorrarekiko eta osagaien portaerarekiko konparazioak direla eta, emaitza nabari gehienak ($AP_{\texttt{id}}$-renak) lehen kasurako lortu dira. Salbuespena, hemen ere, `IZL` hedapena da. Hori ez da Bannarden (2007) hipotesiaren arabera espero genezakeen emaitza (hau da, malgutasunaren neurketa doiagoa eta haztatuagoa lortzen dela osagaien portaerarekiko konparazioa eginez).

VII.16 taulan, erlatiboarekiko malgutasunaren neurketaren emaitzak ditugu. Iruzkin labur batzuk:

- Aurrekoan bezala honetan ere, τ_B-ren balioak t neurriaren azpitik daude.

- $AP_{\texttt{id}}$-ren emaitzak dira, honetan ere, t neurriaren emaitzak gainditzen dituzten bakarrak. Hala ere, ez dira aurreko izen-hedapenen emaitzen mailara iristen.

- Teorian, espero genezake, `erlt_PART` hedapenen detekzioa doitasun handiagokoa denez, horiek bakarrik kontuan hartzeak emaitza hobeak

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
	t neurria	0,197	0,455	0,084	0,383
ERLT	RFR_orok	0,091	0,405	0,132	0,289
	RFR_erltPART_orok	0,053	0,374	0,102	0,283
	KL_div_orok	(-0,090)	0,290	0,047	0,258
	KL_div_erltPART_orok	(-0,052)	0,310	0,050	0,276
	CPMI_orok	0,099	**0,408**	0,122	**0,298**
	CPMI_erltPART_orok	0,052	0,374	0,102	0,283
	Hrel_erlt	0,091	0,405	0,129	0,290
	Hrel_erltPART_orok	0,056	0,374	0,102	0,283
	RFR_izena	(-0,086)	0,274	0,070	0,210
	RFR_aditza	0,097	0,390	0,144	0,266
	RFR_osag	(0,020)	0,323	0,102	0,233
	RFR_erltPART_izena	(-0,089)	0,278	0,065	0,218
	RFR_erltPART_aditza	0,063	0,374	0,130	0,261
	RFR_erltPART_osag	(-0,005)	0,317	0,101	0,228
	KL_div_izena	(-0,133)	0,243	0,057	0,188
	KL_div_aditza	0,097	0,350	0,126	0,242
	KL_div_osag	(-0,022)	0,287	0,086	0,208
	KL_div_erltPART_izena	(-0,144)	0,240	0,060	0,182
	KL_div_erltPART_aditza	(0,039)	0,313	0,100	0,225
	KL_div_erltPART_osag	(-0,071)	0,264	0,074	0,196
	CPMI_izena	(-0,096)	0,277	0,070	0,211
	CPMI_aditza	**0,102**	0,397	**0,149**	0,269
	CPMI_osag	(0,023)	0,338	0,108	0,242
	CPMI_erltPART_izena	(-0,081)	0,286	0,067	0,223
	CPMI_erltPART_aditza	0,068	0,380	0,138	0,261
	CPMI_erltPART_osag	(0,001)	0,337	0,108	0,241

VII.16 Taula: Erlatibodun hedapen-aldakuntzekiko malgutasun morfosintaktikoaren neurketaren emaitzak.

ematea; hala ere, datuek ez dute hori baieztatzen. Bi neurketen arteko aldeak ez dira esanguratsuak. Batetik, horren interpretazio posible bat aurki dezakegu VII.1.3.3 atalean erlatiboaren anbiguotasun-arazoez eman dugun azalpenean: Eustaggerrek perpaus erlatibo jokatuen analisi anbiguoetan egiten dituen akatsak uniformeki banatuta egotea aditzen artean. Bestetik, emaitzetatik atera daitekeen ondorioa da er-

latibo jokatuarekiko eta jokatugabearekiko malgutasunak, oro har, ez direla desberdinak.

- CPMI neurria da, oro har, neurketa onenak eman dituena, baina ez halako alde handiz, batez ere RFRrekin konparatuz gero.

- Konparaziorako erabilitako portaerak direla eta, interesgarria da $AP_{\texttt{id}}$-ren emaitza onena CPMI_aditza neurriak izan duela; gainerako neurketa gehienen kasuan, ikusten da _aditza motako neurketek _izena motakoak baino hobeak direla. Beraz, behatu duguna da konbinazioa idiomatikoagoa izan ahala, erlatiboaren bidezko hedapenak osatzeko gertatzen den malgutasun-galera nabarmenago ageri dela dagokion aditzaren portaerarekin konparatzen badugu.

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
	t neurria	0,197	0,455	0,084	0,383
mugat	SSD_orok	(-0,043)	0,328	0,102	0,237
	KL_div_orok	(-0,020)	0,285	0,087	0,209
	CPMI_orok	0,118	0,395	0,118	0,291
	Hrel	0,119	**0,434**	0,134	0,313
	SSD_izena	(-0,018)	0,340	0,157	0,223
	SSD_aditza	0,055	0,340	0,100	0,254
	SSD_osag	0,064	0,367	**0,161**	0,248
	KL_div_izena	(-0,061)	0,307	0,121	0,208
	KL_div_aditza	(0,015)	0,313	0,084	0,239
	KL_div_osag	(0,006)	0,320	0,120	0,225
	CPMI_izena	0,133	0,425	0,101	**0,331**
	CPMI_aditza	0,090	0,387	0,111	0,290
	CPMI_osag	**0,154**	0,432	0,113	0,329

VII.17 Taula: Mugatasun-aldakuntzekiko malgutasun morfosintaktikoaren neurketaren emaitzak.

Hurrena, mugatasun-aldakuntzekiko malgutasunaren neurketa dugu, eta VII.17 taulan daude emaitzak. Badira aurrekoekiko alde interesgarri batzuk:

- τ_B eta $AP_{\texttt{UF}}$ kontuan hartuta, esan daiteke emaitza orokorrak aurrekoetan baino hobeak direla, t neurriaren mailara iritsi gabe ere. Hala ere, aurrekoetan esapide idiomatikoekin lortutako alde nabariak

banatuagoak daude kasu honetan, kolokazioetan emaitzak hedapenen kasuan baino hobeak baitira. Horrek adierazten digu mugatasun-aldakuntzak ez daudela `id` / `col` bereizketarekin lotuak, UF / ez-UF bereizketarekin baizik.

- Ez da neurri bat besteen gainetik nabarmentzen, baina, bat hautatzekotan, CPMI litzateke eraginkorrena, salbu esapide idiomatikoen kasu berezian, non SSD_osag den neurri onena.

- Konparaziorako erabilitako portaerak direla eta, osagaiekikoa da eraginkorragoa aldakuntza honetan (portaera orokorrarekiko H_{rel} neurriaren balioa ez da osagaiekiko CPMI_osag neurketa baino esanguratsuki hobea).

Azkenik, VII.18 taulan daude ordena-aldakuntzekiko malgutasun morfosintaktikoaren neurketaren emaitzak. Bistakoa da aldakuntza-mota honek ez duela inolako korrelaziorik idiomatikotasunarekin. VII.1.3.2 atalean aurresan genuen emaitza hori, euskara ordena libre samarrekoa delako egitatetik abiatuta.

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
	t neurria	0,197	0,455	0,084	0,383
	SSD_orok	(-0,007)	0,282	0,090	0,202
	KL_div_orok	(-0,003)	0,286	0,087	0,207
	CPMI_orok	(0,021)	0,314	0,083	0,237
	Hrel	(0,023)	0,317	0,081	0,241
	SSD_izena	(-0,149)	0,240	0,060	0,182
	SSD_aditza	(0,023)	0,296	0,078	0,223
ord	SSD_osag	(-0,075)	0,263	0,071	0,195
	KL_div_izena	(-0,117)	0,247	0,063	0,187
	KL_div_aditza	(0,019)	0,293	0,076	0,221
	KL_div_osag	(-0,054)	0,267	0,071	0,199
	CPMI_izena	(0,033)	0,311	0,084	0,234
	CPMI_aditza	**0,096**	**0,352**	**0,096**	**0,261**
	CPMI_osag	0,068	0,334	0,090	0,250

VII.18 Taula: Ordena-aldakuntzekiko malgutasun morfosintaktikoaren neurketaren emaitzak.

Azkenik, gogoraraztekoa da malgutasun morfosintaktikoa, idiomatikotasunaren beste propietateak ez bezala, magnitude konplexua dela, deskribatu ditugun zenbait aldakuntza-motarekiko malgutasun bakunen konbinazioa litzatekeena. Horien neurketen emaitzak neurri bakarrean konbinatzea da hurrengo urrats logikoa. Baina ikusi dugu erabilitako neurketa batzuetan, batez ere malgutasunaren noranzkoaren alderditik, neurketak ezin direla konbinatu. Beraz, malgutasunaren neurketa bateratu bat egin dugu ahal izan den bi neurri hauen kasuan: KL dibergentzia eta CPMI. VII.19 taulan, izenaren `DET`, `ADJ` eta `IZL` hedapenekiko malgutasunen baturak (_hedap) eta aldakuntza guztiekiko neurketen baturak (_big) erabiliz lortutako CPMI neurriaren araberako emaitzak bistaratu ditugu.

Taula horretako balioek ez dituzte gainditzen aurreko tauletan emandako malgutasun bakunen emaitza onenak. Dena den, VII.2 kapituluan izango dugu aukera, ikasketa automatikoaren testuinguruan, malgutasun morfosintaktikoaren osagaiak modu integratuan erabiltzeko.

neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
ausaz	0,000	0,309	0,070	0,234
t neurria	0,197	0,455	0,084	0,383
CPMI_hedap_orok	0,074	0,400	0,175	0,257
CPMI_big_orok	0,108	0,415	0,141	0,292
CPMI_hedap_izena	0,013	0,318	0,118	0,220
CPMI_hedap_aditza	0,077	0,391	0,135	0,275
CPMI_hedap_osag	0,072	0,373	0,137	0,255
CPMI_big_izena	0,077	0,365	0,106	0,269
CPMI_big_aditza	0,115	0,412	0,135	0,295
CPMI_big_osag	0,131	0,418	0,130	0,304

VII.19 Taula: Izenaren `DET`, `ADJ` eta `IZL` hedapenekiko malgutasunen baturak (_hedap) eta aldakuntza guztiekiko neurketen baturak (_big) erabiliz lortutako CPMI neurriaren araberako emaitzak.

VII.1.4 Malgutasun lexikalaren neurketa, ordezkagarritasunaren bidez

II.1.2.3 atalean azaldu dugunez, UF batzuek murrizketa lexikala agertzen dute, hau da, osagaietako bat edo biak ezin dira haien sinonimoez, kuasisinonimoez edo semantikoki erlazionatutako hitzez ordezkatu, UFaren ezaugarriak modifikatu gabe (instituzionalizazioa, esanahia, idiosinkrasia). III.9

atalean ikusi dugu ezaugarri hori kuantifikatzeko teknikaren oinarria konbinazioaren osagaien ordezkagarritasuna neurtzea dela.

VII.1.4.1 Baliabideak

Bigramen osagaien ordezkagarritasuna neurtzeko, lehenik eta behin, osagaien ordezkoak behar ditugu. Horretarako, euskarazko bi baliabide hauetatik abiatu gara:

- *Sinonimoen Kutxa*, Elhuyar Fundazioak argitaratutako sinonimo-hiztegia[14].
- Euskal WordNet[15] (Pociello et al., 2011). Ikerketa honetan, 3.0 bertsioa erabili dugu.

Baliabide bakoitzeko sarrerak eta *synset*ak prozesatu ditugu, sinonimo-bikoteak sortzeko. VII.20 taulan, bi iturri horietako izen- eta aditz-kategoriako sinonimo-bikoteen kopuruak eman ditugu.

kategoria	ELH_SK	EusWN 3.0
IZE	29 597	38 480
iz.	28 038	
adj./iz.	141	
iz./izlag.	71	
iz./izond.	1 348	
ADI	9 566	14 216
Totala	**39 163**	**52 696**

VII.20 Taula: Elhuyarren *Sinonimoen Kutxa*tik eta Ixa taldearen Euskal WordNet 3.0-tik eratutako izen- eta aditz-kategoriako sinonimo-bikoteen kopurua, hitz bakunak eta marradun izen-elkarteak erabiliz, eta adierabereizketa kontuan hartuta.

Taulako datuak ondo interpretatzeko, zehaztapen hauek egin behar dira:

- Elhuyarren *Sinonimoen Kutxa*ren kasuan, kategoria bikoitza duten sinonimoak ere kontuan hartu dira izenaren ordezkoak lortzeko (*adj./iz.*, *iz./izlag.* eta *iz./izond.*).

[14] *Sinonimoen Kutxa*. 2010. 3. ed. Elhuyar Fundazioa. Usurbil.
[15] http://ixa2.si.ehu.es/mcr/

- Hitz bakunak diren sinonimoak bakarrik hartu ditugu kontuan, edo, izenen sinonimoen kasuan, marraz lotutako izen-elkarteak. Arrazoia da konbinazio "ordezkatua" bilatzeko espazioa V.3.1 atalean deskribatu dugun bigrama-erauzketa baten emaitza dela, eta hitz bakunez osatuak dira, edo, izenen kasuan, lema bakarrean etiketatutako marraz lotutako izen-elkarteak. Zehazki, $w = \pm 1$ eta $f \geqslant 3$ erauzketa erabili dugu.

- Bi baliabideetan, eta batez ere `EusWN`en, sinonimo-bikote bera zenbait adiera edo *synset*etan gertatzen da. Taulako datuak adiera-bereizketa kontuan hartuta ateratakoak dira, hau da, adierak kolapsatu gabe.

Esperimentuen diseinuaren atariko urrats batzuk egin ditugu baliabide bakoitza bereiz erabiliz, eta emaitzek erakutsi dute maiz bigramaren osagaietako batek, edo biek, ez dutela sinonimorik hiztegian, edo, izanda ere, ordezko sinonimo horrekin osatutako konbinazioak ez duela agerpenik corpusean. Horrelakoetan, zaila da egoera interpretatzea. Hitz batek sinonimorik ez badu, badirudi ez duela zentzurik malgutasun lexikalaz hitz egitea; baina baliabide lexikalarekiko mendekotasuna ere agerikoa da. Sinonimoak izanda ere, nola interpretatu corpusean ez izatea horiekin osatutako konbinaziorik? Erabateko murrizketa lexikalaz hitz egin genezake? Oraingoan, corpusarekiko mendekotasuna nabarmentzen da. Adierazle batzuk baditugu, beraz, metodo hau sentikorra dela esateko esperimentuetan erabiltzen ditugun baliabideen estaldurarekiko. Nolanahi ere, erabaki egin beharra dago horrelakoetan zein balio hartuko duen malgutasuna neurtzeko erabiltzen dugun neurriak (ikus VII.1.4.2 atala).

Bestetik, ordezkagarritasuna neurtzeko parametro bat da hiztegiko adiera-bereizketa kontuan hartu ala ez. Teorikoki, ordezkagarritasuna esanahi bereko hitzez osatutako konbinazioen artean egiaztatu behar da. Lehen arazoa da ez dugula zehazterik zein diren aztergai dugun konbinazioaren osagai bakoitzaren adierak, edo, behintzat, osagai bakoitzari zein adiera dagokion ordezkoen iturritzat erabiliko dugun hiztegian. Ordezkagarritasuna oso handia bada hiztegiko adiera bateko sinonimoekin baina oso baxua beste adiera batekoekin, nola interpretatu emaitza hori? Pentsa genezake gure bigrama ez dela idiomatikoa lehen adieran, eta bai bigarrenean. Bi interpretazio horietako bat ausaz edo arbitrarioki hautatzeak eragin handia izan dezake karakterizazioan.

Bi arazo horiek pentsarazi digute baliabideen estaldura handitzea komeni dela, osagaien ordezkoen konbinazioen agerpenik ezeko kasuak albait gutxitzearren, nahiz eta jakitun izan hori eginez ezin izango dela adiera-bereizketa kontuan izan, baliabide bakoitzaren adiera-banaketa desberdina

delako, eta haien arteko mapatzea egitea ahalmenetik kanpo delako. Horretara, aipatu bi baliabideak kondizio horietan konbinatu ditugu, eta ELHWN izena eman diogu emaitzari.

Estaldura handitzeko beste saio bat izan da bigramaren osagai bakoitzaren sinonimoak "hedatzea", EusWNen dituen ahaideak (*siblings*, edo hiperonimo bereko hitzak) gehituz. ELHWNhedap izena du hedapen horrek.

VII.21 taulan, estaldura handitzeko bi saio horien emaitzak ageri dira, ELH_SK eta EusWN baliabideen sinonimo-bikoteen kopuruen ondoan, adiera-bereizketa kontuan hartu gabe, denak ere.

kategoria	ELH_SK	EusWN 3.0	ELHWN	WNhedap	ELHWNhedap
IZE	29 407	29 603	52 434	548 454	570 404
ADI	9 333	7 898	15 061	87 567	94 371
Totala	**38 740**	**37 669**	**67 495**	**636 021**	**664 775**

VII.21 Taula: Adiera-bereizketa kontuan hartu gabe eratutako sinonimo-bikoteen bost bildumak.

Azkenik, Van de Cruys eta Moiróni (2007) jarraituz, bigramen osagaien thesaurus distribuzional bat osatu dugu corpusetik, semantikoki antzekoenak diren ordezkoak eskuratzeko (kategoria berekoak, betiere). Lan hori egiteko, Saralegi et al.-en (2008) lanean garatutako tresna egokitu dugu.

Thesaurusa osatzeko, osagai bakoitzaren testuinguru-bektoreak osatu ditugu (eduki-hitzak erabiliz: IZE, ADI eta ADJ), bektoreen pisuetan egiantz-arrazoia erabiliz, eta bektoreen arteko antzekotasun-neurritzat kosinua erabiliz. EB_antzdist izena eman diogu (Egunk+Berria corpusetik eratutako thesaurus distribuzionala). Esperimentuetan, glosarioko hitz bakoitzaren lehen 20 antzekoenak hartu ditugu kontuan osagaien ordezkotzat erabiltzeko (EB_antzdist_20).

Prozedura honek badu antza Linek (1999) eratutako thesaurusa erabiltzearekin. Nolanahi ere, guk ezin izan dugu horrelakorik planteatu; batetik, ikerlan horietan N_OBJ+V konbinazioak aztertzen dira, eta guk esparru zabalagoko izena+aditza konbinazioak hartu ditugu aztergaitzat; baina, batez ere, arazoa da inguru esperimentala diseinatu dugunean ez dugula modurik izan SUBJ, OBJ eta gainerako funtzio sintaktikoak analizatuko zituen euskarazko prozesatzailerik modu masiboan eta doitasun handiz erabiltzeko. Etorkizuneko lantzat kontsideratzen dugu aukera hori.

VII.22 taulan, *adostasun* izenaren 20 izen antzekoenak bildu ditugu.

osagaia	kat.	antzekoa	ord.	cos
adostasun	n	garaipen	1	0,84282
adostasun	n	kontsentsu	2	0,83784
adostasun	n	akordio	3	0,83587
adostasun	n	berdintze	4	0,76968
adostasun	n	balentria	5	0,66952
adostasun	n	helburu	6	0,66817
adostasun	n	arrakasta	7	0,63112
adostasun	n	erreskan	8	0,61732
adostasun	n	berdinketa	9	0,60867
adostasun	n	eserleku	10	0,60645
adostasun	n	aurreakordio	11	0,60421
adostasun	n	saskiraketa	12	0,60172
adostasun	n	emaitza	13	0,59770
adostasun	n	esker	14	0,59041
adostasun	n	eskainu	15	0,58304
adostasun	n	sailkatze	16	0,57603
adostasun	n	gehiengo	17	0,56599
adostasun	n	puntu	18	0,56076
adostasun	n	inkulpazio	19	0,55724
adostasun	n	quorum	20	0,54761

VII.22 Taula: *Adostasun* hitzaren antzekotasun distribuzional handieneko hitzak.

Ordezkagarritasuna neurtzeko baliabideen estalduraren ideia bat iradokitzeko, VII.23 taulan eman ditugu osagairen baten kasuan ordezkoarekin osatutako konbinazio-aldaerarik aurkitu ez dugun kasuen kopuruak, dela baliabidean sinonimorik ez duelako, dela konbinazioaren aldaerak corpusean agerpenik ez duelako. Agerikoa da ez direla kopuru txikiak, eta hori bide esperimental honen handicap bat izan daiteke.

VII.1.4.2 Neurriak

III.9 atalean aurkeztutako neurketa-metodoetatik, azken urteetan garatutako bi aukeratu ditugu karakterizatu nahi dugun konbinazioa eta haien osagaien ordezkoez osatutako konbinazioak konparatzeko:

Baliabidea	ordezkorik ez
ELH_SK	674
EusWN 3.0	552
ELHWN	475
WNhedap	299
ELHWNhedap	238
EBantzdistr_20	207

VII.23 Taula: Ebaluazio-erreferentziako 1 145 bigrametatik, gutxienez osagai baten ordezko konbinaziorik ez duten edo corpusean ordezkoaren agerpenik ez duten bigramen kopurua, ordezkoak aurkitzeko erabilitako baliabidearen arabera.

- Van de Cruys eta Moirónen (2007) R indizea, bigramaren eta ordezko aldaeren arteko KL dibergentzian oinarritua. R_{nv} eta R_{vn} kalkulatu ditugu, hurrenez hurren, aditz batek izen jakin batekiko duen joera eta izen batek aditz batekiko duena adierazteko, bere ordezkoek izen edo aditz horiekiko duten joerarekin konparatuta. Gogoan izan behar dugu konbinazioak ordezkorik ez duenean, edo ordezkoak corpusean agerpenik ez duenean, ordezkoaren $p(n|v) = 0$ dela, eta egileek 1 balioa esleitzen diotela R indizeari. Baina, egileen arabera, R-ren balio-tartea $[0, +1]$ da, eta, ondorioz, ordezkoen daturik ez dugun kasuetan, konbinazioa lexikalki erabat finkatua dela asumitzen dugu. Horrek zarata sor dezake, zeren konbinazio libre asko rankingaren lehen posizioetara eraman baititzake. VII.23 taulako datuak ikusita, bistan da ordezkorik ezeko egoera ez dela `id` edo `col` kategorietakoen arazoa soilik, eta zuhurragoa da horrelakoetan R-ri 0 balioa esleitzea.

- Fazly eta Stevensonen (2007) $\text{Fixedness}_{\text{lex}}$ neurria, bigramaren eta haren osagai bakoitza ordezkatuz sortutako aldaeren MI balioen arteko z neurria dena. Egileek ez dute bereizten izenaren eta aditzaren ordezkagarritasunen artean, baterako kalkula egiten baitute. Baterako neurketaz gain (z_PMI_NV), z_PMI_V eta z_PMI_N ordezkagarritasunak bereiz kalkulatu ditugu. Ordezkorik ezean, egileek ez dute zehazten balioa esleitzeko irizpidea; dena den, kasu horretan z neurria zehaztugabea da, eta 0 balioa esleitu diogu. Hori da irtenbiderik zentzuzkoena, kontuan izanik z neurriaren balio-tartea $[-\infty, +\infty]$ dela.

Beraz, R_{nv} eta z_PMI_V neurriek konbinazio baten aditzaren ordezkagarritasuna neurtzen dute; R_{vn} eta z_PMI_N neurriek, berriz, izenarena.

Neurri horien kalkulua egiteko programazioa Perl lengoaian garatu dugu.

VII.1.4.3 Emaitzak

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
	t neurria	0,197	0,455	0,084	0,383
	R_{nv}	(0,091)	0,358	0,082	0,281
	R_{vn}	(0,065)	0,354	0,081	0,280
ELHWN	z_PMI_V	(0,089)	0,349	0,083	0,270
	z_PMI_N	(0,038)	0,330	0,064	0,272
	z_PMI_NV	0,079	0,339	0,072	0,272
	R_{nv}	0,074	0,361	0,081	0,287
	R_{vn}	(0,029)	0,339	0,102	0,246
WNhedap	z_PMI_V	**0,110**	0,365	0,120	0,260
	z_PMI_N	(0,083)	0,356	0,073	0,290
	z_PMI_NV	0,095	0,348	0,092	0,263
	R_{nv}	0,063	0,346	0,083	0,268
	R_{vn}	(0,020)	0,332	0,100	0,241
ELHWNhedap	z_PMI_V	0,101	0,357	**0,122**	0,250
	z_PMI_N	0,084	0,355	0,072	0,290
	z_PMI_NV	0,094	0,349	0,096	0,260
	R_{nv}	0,066	**0,381**	0,066	**0,323**
	R_{vn}	(-0,032)	0,311	0,081	0,236
EBantzdistr_20	z_PMI_V	(0,045)	0,344	0,091	0,266
	z_PMI_N	(-0,026)	0,295	0,072	0,226
	z_PMI_NV	(0,014)	0,321	0,094	0,236

VII.24 Taula: Malgutasun lexikalaren neurrien emaitzak.

VII.24 taulan eman ditugu malgutasun lexikala neurtzeko esperimentuen emaitzak. Hauek dira nabarmentzeko alderdiak:

- Emaitzak aurreko hiru propietateenak baino apalagoak dira. AMekin konparatuz, $AP_{\texttt{id}}$-ren emaitza batzuetan bakarrik gainditzen da t

neurria: eta alde txikiz, oso urruti DSim neurrietatik, baita MS neurrietatik ere.

- WordNeteko hedapenaren bidez aberastutako baliabideek (`WNhedap` eta `ELHWNhedap`) hobekuntza arin bat dakarte hedapenik gabekoekiko (`ELHWN`), zehazki esapide idiomatikoen z_PMI_V neurriaren emaitzetan (ezen ez kolokazioetan); dena den, jauzia txikiegia da ondorio sendorik ateratzeko.

- $AP_{\texttt{UF}}$ eta $AP_{\texttt{col}}$-en emaitza onenak `EBantzdistr_20` baliabidearekin lortu dira. Bada besteekiko alde nabarmenago bat aurreko atalean aipatu duguna baino; hala ere, $AP_{\texttt{id}}$-ren emaitzak ausazko ranking baten mailakoak dira.

- Emaitza onenak, ia beti, aditza ordezkatuz egindako esperimentuetan lortu dira (R_{nv} edo z_PMI_V neurriak).

VII.1.5 Esperimentu bakunen emaitzen analisia

Aurreko ataletan aurkeztu ditugun esperimentu bakunen emaitzak globalki aztertuz, analisi osoago bat egin eta lehen ondorio batzuk atera daitezke. Horien azalpenaren lagungarri, VII.25 taulan laburbildu ditugu idiomatikotasunaren neurketa bakunen emaitza nabarmenak.

1 Antzekotasun distribuzionaleko neurriak (DSim) dira idiomatikotasun-mailarekin korrelazio onena dutenak. Kendall τ_B-ren emaitzetan alde handia dago DSim esperimentuen eta gainerakoen artean. Horrek adierazten du konposizionaltasun-maila gradualki handituz doala esapide idiomatikoetatik kolokazioetan barrena konbinazio libreetaraino doan continuumean; hau da, ez dela esapide idiomatikoen ezaugarri esklusiboa. Agerkidetza-esperimentuak dira hurrenak (AM), malgutasun morfosintaktikokoak gero (MSFlex), eta malgutasun lexikalekoak azkenak (LFlex).

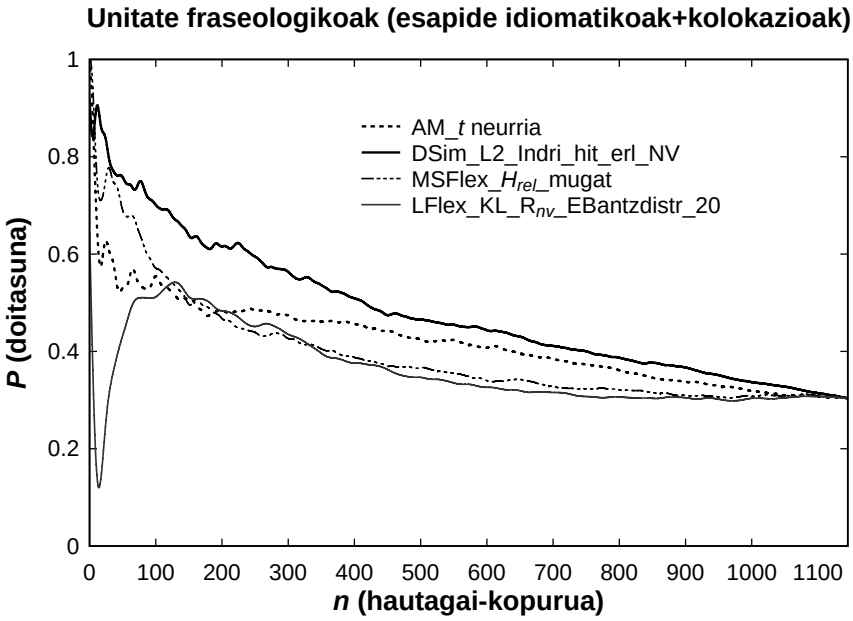
2 UFen erauzketan (UF-kategoria kontuan hartu gabe), $AP_{\texttt{UF}}$-ren balioak aurreko puntuko ondorioekin koherenteak dira: `id-col-free` ordenazio onena eratzeaz gain, UFak (`id` eta `col` kategoriak bereizi gabe) doitasun handienaz erauzten dituzte DSim neurriek. Grafikoki ikus dezakegu hori VII.6 irudiko P kurbetan (idiomatikotasunaren propietate bakoitzetik, emaitza onenak izan dituen neurria eman dugu). Bestetik,

	neurria	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
	ausaz	0,000	0,309	0,070	0,234
AM	t neurria	0,197	0,455	0,084	0,383
	χ^2	(-0,037)	0,302	0,119	0,206
DSim	L2_Indri_hit_erl_NV	**0,322**	**0,566**	0,294	0,354
	L2_KL_hit_erl_NV	0,309	0,551	**0,320**	0,332
	L2_Indri_rankV_hazt	0,320	0,552	0,130	**0,431**
MSFlex	CPMI_osag_mugat	0,154	0,432	0,113	0,329
	H_{rel}_mugat	0,119	0,434	0,134	0,313
	CPMI_orok_ADJ	(0,045)	0,387	0,202	0,237
	CPMI_izena_mugat	0,132	0,425	0,101	0,331
LFlex	z_PMI_V_WNhedap	0,110	0,365	0,120	0,259
	R_{nv}_EBantzdistr_20	0,066	0,381	0,066	0,323
	z_PMI_V_ELHWNhedap	0,108	0,357	0,122	0,250

VII.25 Taula: Idiomatikotasunaren lau osagaiak neurtzeko egindako esperimentu bakunen emaitza onenen laburpena. Kendall τ_B koefizientea (idiomatikotasun-ranking ideal batekiko hein-korrelazioa); eta batez besteko doitasunak (AP), UFen rankingerako ($AP_{\texttt{UF}}$), oro har, eta AP espezifikoak esapide idiomatikoetarako eta kolokazioetarako.

$n \lesssim 150$ tartean, MSFlex neurriak t neurriak baino P handiagoa lortzen du. Hortik aurrera, ordea, portaera txarragoa du, eta horregatik da $AP_{\texttt{UF}}$-ren balioa apalagoa.

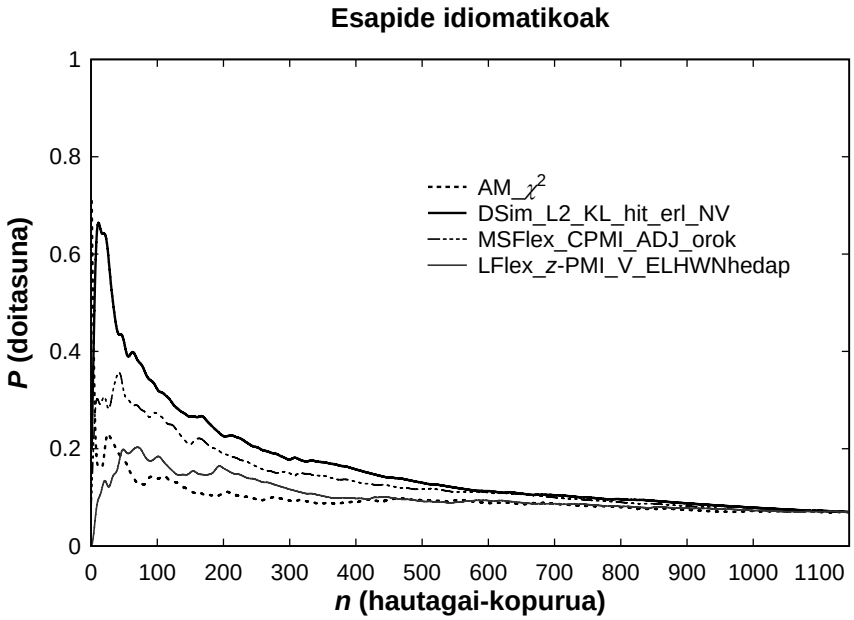
3 DSim neurri onenak testuinguru-dokumentuen arteko antzekotasuna neurtzen duten Indri indizea eta KL dibergentzia dira, L2 esperimentu-modalitatean, ohiko WSM ereduko kosinuaren gainetik. Emaitza hau garrantzitsua da, IR tresna batek egiteko honetan izan dezakeen aplikazioaren agergarria delako; eta interesgarria, zeren L2 modalitatean bigramaren agerpen bakoitza ebaluatzen baita osagaien testuinguruen bildumaren kontra, eta gero portaera orokorra batez bestekoaren bidez karakterizatu. Horrek iradokitzen du etorkizunean metodo baliagarria izan daitekeela konbinazio baten agerpenak banan-banan ebaluatu, eta izan litezkeen interpretazio desberdinak, idiomatikoak edo literalak, bereizteko.



VII.6 Irudia: UFen idiomatikotasun-rankingen doitasun-kurbak. Idiomatikotasunaren lau propietateetatik, neurketan emaitza onena izan duen neurri bana.

4 Esapide idiomatikoen erauzketan nabari da batez ere DSim neurrien gailentasuna, $AP_{\texttt{id}}$-ren balioek argi erakutsi dutenez. Esapide idiomatikoak konposizionaltasun txikieneko UFak direnez, emaitza hori espero izatekoa zen, eta berretsi egiten du konposizionaltasun semantikoa karakterizatzeko antzekotasun distribuzionala eraginkorragoa dela agerkidetza-neurriak baino. VII.7 irudian, esapide idiomatikoen P kurbak daude.

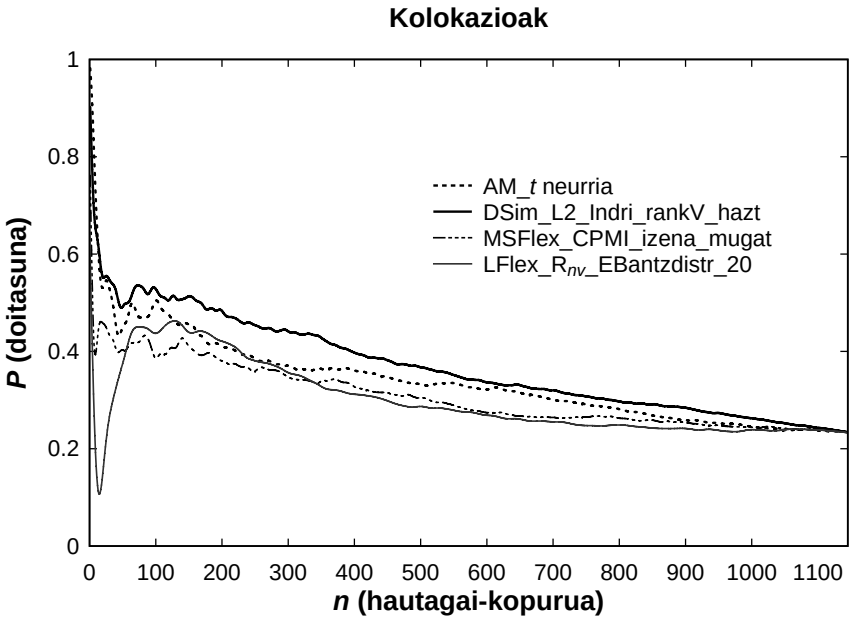
5 Kolokazio-erauzketan, AM neurri onena (t neurria) DSim neurrien lehiakide handia da, baina, eskuarki kolokazioen ezaugarri aipatuenetakoa idiosinkrasia estatistikoa bada ere, emaitza onena aditzaren semantika neurtzen duen Indri indize batek izan du, aditzak izenarekin konbinazioan dituen testuinguruak izenik gabe agertzen den testuinguruekin konparatzen dituen batek, hain zuzen ere (L2_Indri_rankV_hazt). $n > 120$ puntutik aurrera da hori nabaria, VII.8 irudiko kolokazioen P kurbek erakusten dutenez. Emaitza hori bat dator kolokazioak erdikon-



VII.7 Irudia: Esapide idiomatikoen idiomatikotasun-rankingen doitasun-kurbak. Idiomatikotasunaren lau propietateetatik, neurketan emaitza onena izan duen neurri bana.

posizionalak direlako ikuspegiarekin, eta Venkatapathy eta Joshiren (2005) emaitzekin.

6 MSFlex: emaitza onenak, oro har, mugatasun-aldakuntzen neurketek lortzen dituzte. Dena den, horiek ez dituzte agerkidetza-esperimentuen emaitzak gainditzen, $AP_{\texttt{id}}$ kasuan izan ezik (H_{rel}_mugat), eta esanguratsua ez den tarteagatik. Horrek adierazten du mugatzailearekiko finkotasuna nabarixeagoa dela esapide idiomatikoetan, baina diskriminatze-ahalmena ez dela oso handia. Kasu batean bakarrik dira AMenak baino nabarmenki hobeak, `ADJ` hedapenaren $AP_{\texttt{id}}$-ren balioan hain zuzen ere. Oso argi ikus dezakegu hori VII.7 irudian. Azkenik, bistan da MSFlex neurketen $AP_{\texttt{col}}$-en emaitzak balio apalenen artean daudela. Taulako neurketetan, gainera, ez dugu ikusten hori $AP_{\texttt{id}}$ altu batekin batera gertatzen denik; beraz, gure interpretazioa da kolokazioen ezaugarrien artean malgutasun morfosintaktikoa dela bereizgarritasun txikienekoa. Hori bat letorke fraseologia teorikoak

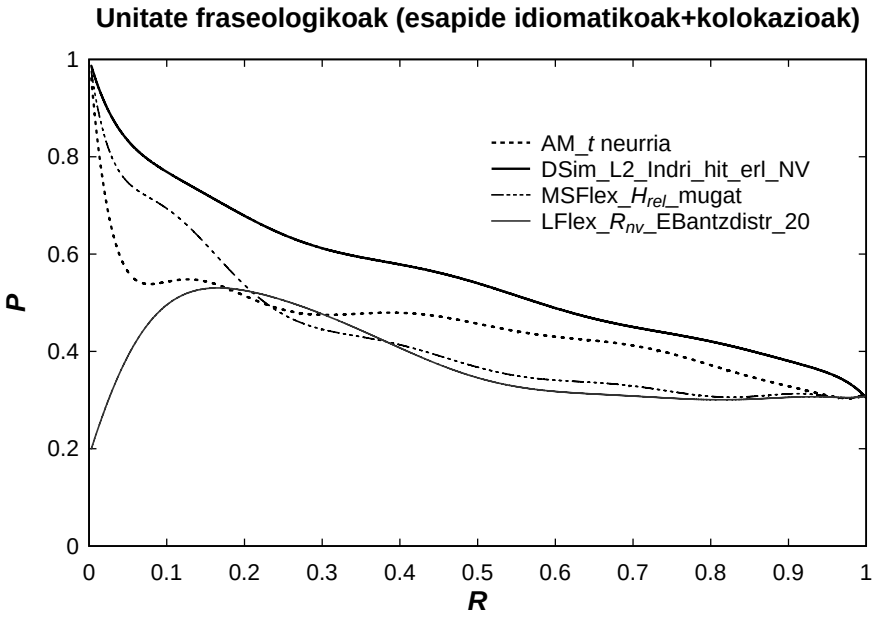


VII.8 Irudia: Kolokazioen idiomatikotasun-rankingen doitasun-kurbak. Idiomatikotasunaren lau propietateetatik, neurketan emaitza onena izan duen neurri bana.

esandakoarekin (ikus II.2.2.4 atala).

7 LFlex: emaitzak espero baino txarragoak dira. Esapide idiomatikoen erauzketan, AMen mailako AP emaitzak dituzte, eta, kolokazio-erauzketan, MSFlex-enak baino apalagoak, batez ere $n < 100$ tartean, non konbinazio libre asko sartzen baitira (ikus VII.9 eta VII.11 irudiak). Esapide idiomatikoen zein kolokazioen definizioa egin dugunean, azaldu dugu murriztapen lexikala giltzarritzat jo ohi dela, eta garrantzi hori ez da emaitzetan islatzen. Espero gabeko emaitza eskasak izan ditugu propietate hau neurtzeko esperimentuetan, eta esplikazioak bilatzea ezinbestekoa da (ikus VIII.1 atala).

Everten iritziz (Evert, 2005: 143), emaitza aurkezteko modu intuitiboena P/R grafikoa da, neurrien arteko aldeak nabarmenago bistaratzen baitira. VII.9, VII.10 eta VII.11 irudietan, hurrenez hurren, UF, esapide idiomatikoen eta kolokazioen P/R kurbak daude. Hemen ere, idiomatikotasunaren propietate bakoitzetik, emaitza onenak izan dituen neurria eman dugu.

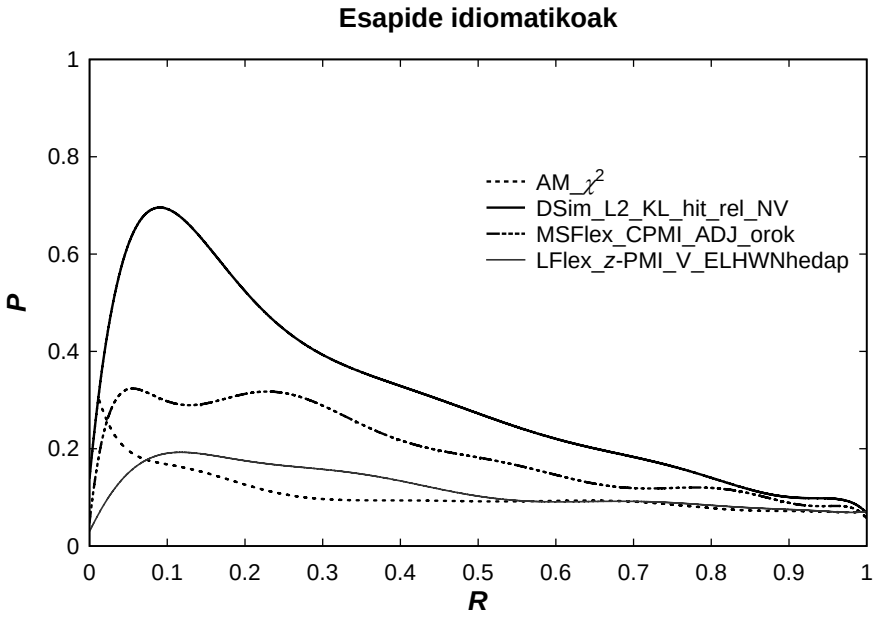


VII.9 Irudia: UFen idiomatikotasun-rankingen P/R kurbak.

UFen kasuan (VII.9 irudia), oso nabaria da DSim neurriaren gailentasuna P/R-ren balio guztietan barrena. Oro har, kurbaren profilak badu antza Evertek erakutsitako emaitzen profilarekin. Izan ere, batean zein bestean, doitasuna nahiko azkar jaisten da estaldurarekin. Horrek adierazten digu, estaldura handia behar den aplikazioetan, ezin izango dugula doitasun handirik espero. Esaterako, UFen % 70 erauzi nahi baditugu, 0,5eko doitasuna izango dugu, hau da, erauzitakoen erdiak ez dira UF izango.

Esapide idiomatikoak direla eta (VII.10 irudia), kurbak agerian uzten du DSim neurriak besteen oso gainetik dabiltzala. Puntako doitasuna (0,7) 0,1eko estaldurarako lortzen da, esapide idiomatikoen % 10 eskuratzen denean. 0,3ko estalduratik aurrera, AM neurriak ez du ausazko rankingari dagokion oinarri-lerroa gainditzen; orobat LFlex neurriak, 0,5eko estalduratik aurrera.

Azkenik, kolokazioen P/R kurbak (VII.11 irudia) nabariago erakusten du VII.8 irudiak iradokitakoa. Aditzarekiko antzekotasun semantikoa neurtzen duen DSim neurria da onena, baina 0,17ko estalduratik aurrera joan behar da hori nabaritzeko.



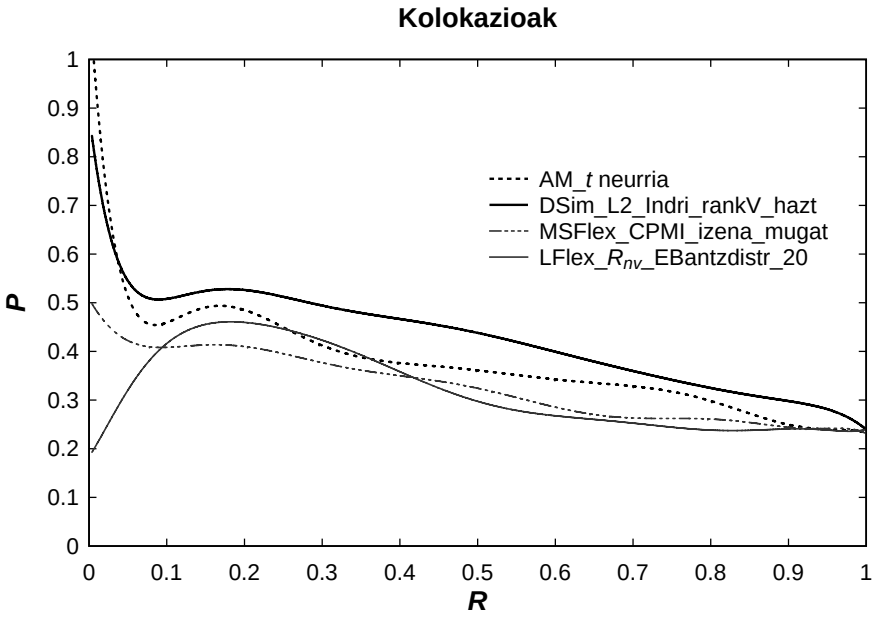
VII.10 Irudia: Esapide idiomatikoen idiomatikotasun-rankingen P/R kurbak.

Emaitzen azterketa konparatibo honek ikasketa automatikotik espero dezakegunaren lehen ideia bat eman du.

VII.2 Propietateen integrazioa: ikasketa automatikoa

Aurreko esperimentuetan, idiomatikotasunaren propietateen banakako neurketa egin dugu, eta ranking-ataza batera orientatu ditugu esperimentuak. Emaitza horien ebaluazioak erakutsi digu neurketa-metodo bakoitzak sortzen duen rankingak edo ordenazioak zenbaterainoko korrelazioa duen idiomatikotasunaren continuumarekin. Halaber, metodo bakoitzak esapide idiomatikoak eta kolokazioak erauztean lortzen duen doitasuna ere agertu dugu.

Idiomatikotasuna fenomeno konplexua izanik, helburuen atalean (I.3 atala) planteatutako hirugarren ikergaiaren erantzuna bilatzea da atal honetan aurkeztuko ditugun esperimentuen helburua. Orain arte propietate bakoitzaz ikasi duguna konbinatzea da ideia, eta, horretarako, ikasketa automatikoaren bidezko sailkatze-ataza eraman dugu aurrera. Gure asmoa ez da ikasketa automatikoaren arloko ikerketa egitea izan, arlo horretako teknika



VII.11 Irudia: Kolokazioen idiomatikotasun-rankingen P/Rkurbak.

estandarrak ataza honetan aplikatzea, eta UFen karakterizazioari dagozkion ondorioak ateratzea baizik.

VII.2.1 Esperimentuen diseinua

Ikasketa automatikoko esperimentuak egiteko, Weka paketea erabili dugu (Hall et al., 2009). Datu-meatzaritzako atazetarako erabiltzen diren ikasketa automatikoko algoritmoen multzoa da[16]. Hainbat algoritmorekin aurreprobak egin ditugu, esperimentuetan erabiltzekoak aukeratzeko. Azkenean, sei metodo hauek erabili ditugu, mota ugaritakoak eta literaturan kontrastatuak direnak:

- *Naive Bayes*. Adibide bat klase batean sailkatzeko, haren probabilitatea maximizatzen duen klasea aukeratzen da, Bayesen teoreman oinarrituta. Atributuak independenteak direla suposatzen du, eta, maiz, baldintza hori ez da betetzen, baina dakarren sinplifikazioa praktikoa da atributu asko daudenean, edo atributu guztien arteko konbinazioen

[16] http://www.cs.waikato.ac.nz/ml/weka/

probabilitateak kalkulatzeko nahikoa adibide ez dagoenean (Manning eta Schütze, 1999: 607). Sarritan, oinarri-lerrotzat hartzen da (Rennie et al., 2003).

- J4.8. Erabaki-zuhaitzen C4.5 algoritmoaren inplementazioa. Erabaki-zuhaitzak "zatitu eta irabazi" algoritmoan oinarrituak dira. Lehendabizi, atributu bat aukeratzen da erro-adabegitzat, eta horren balio adina adar sortu. Hortik aurrera, errekurtsiboki, adar bakoitzerako errepikatzen da prozesua. Beraz, *greedy* moduko algoritmoa da, eta giltzarri da une bakoitzean eta, batez ere, hasieran zer atributu aukeratzen den (Arrieta, 2010: 22-23). HPan, ia ataza guztietan erabiltzen dira erabaki-zuhaitzak.

- *Random Forest*. Erabaki-zuhaitzak ausaz hautatuz eratutako algoritmoa (Breiman, 2001). Hainbat zuhaitz sortzen dira, baina, zatiketa egiteko unean, adabegi bakoitzerako aldagai-multzo mugatu bat hartzen da kontuan, eta multzo hori ausaz hautatzen da; azken sailkatzaileak erabaki-zuhaitz guztiez osatua da, eta erabakia botazioz hartzen da (Pérez, 2006: 45). HParen arloan, HAUen sailkapenean erabili da (Gayen eta Sarkar, 2013; Kontonatsios et al., 2013), eta definizio-erauzketan, besteak beste (Kobyliński eta Przepiórkowski, 2008).

- PART. Erregela bidezko algoritmoa. Erregelak inferitzeko, erabaki-zuhaitz partzialak sortzen dira behin eta berriz, eta bi paradigma onenak erabiltzen dira, "zatitu eta irabazi" ikaste-teknika aplikatuz (Kotsiantis, 2007: 254). HP arloan ez da oso erabilia, baina badira adibideak entitate-erauzketan (Marrero et al., 2009) eta itzulpen automatikoan (Specia, 2005).

- SMO (*Sequential Minimal Optimization*). Funtzio linealetan oinarritua den *Support Vector Machine* edo sostengu-bektoreen makinen modalitatearen inplementazio bat. Bi datu-multzo izaki, marjina handieneko hiperplanoa bilatzen saiatzen da. Hiperplano hori da alde bateko eta besteko instantziak elkarrengandik urrutien jartzen dituena. Hiperplanotik gertuen dauden bektoreei *sostengu-bektore* deritze. Instantziak linealki banagarriak ez direnean, prozedura hori *kernel* deituriko funtzioen bidez orokortu daiteke. SVM algoritmoak atributu asko dituzunean dira bereziki eraginkorrak, eta zarata kudeatzeko ahalmen handia dute; bestetik, HPan oso erabiliak dira gaur egun (Arrieta, 2010: 25-26).

- *Logistic Regression* (erregresio logistikoa). Klase bakoitzaren probabilitatea funtzio logistiko bat erabiliz estimatzen du, eta pisuak egiantza maximizatze aldera hautatzen ditu (Witten eta Frank, 2005: 121-122). HPko arlo batzuetan erabili da, hala nola entitate-erauzketan (Lin eta Wu, 2009) eta adiera-desanbiguazioan (Vickrey et al., 2005).

Atributu gisa, idiomatikotasunaren propietateen banakako esperimentuen emaitzak erabili ditugu (VII.1.1.1, VII.1.2.5, VII.1.3.5 eta VII.1.4.3 ataletan erakutsitako emaitzak). Bi neurri-familia hauek ez ditugu kontuan hartu, emaitza nabarmen txarragoak izan direlako: WSM sortako JSdiv, eta Infomap. Ondorioz, hau da propietateen araberako atributu-kopurua: AM 7; DSim 40; MSFlex 34; eta LFlex 20.

Neurketa horietan, hainbat neurri erabili dira, eta esperimentu-modalitateak ere ugariak izan dira. Lehen hurbilketa batean, aukera bat izan daiteke informazio hori dena erabiltzea, eta, ikasketa-prozesuan barrena, emaitzek atributu eraginkorrenak zein diren erakustea. Azken buruan, gure helburua da sailkatzailearen errendimendua optimizatuko duen atributu-multzoa zehaztea, gerora aplikazio praktikoetarako gara litezkeen tresnetan horiei dagozkien neurrien kalkulua soilik inplementatzeko.

Dena den, aipatu lau atributu-familia horiez gain, eta Fazly eta Stevensoni (2007) jarraituz, konbinazioaren aditza ere sartu dugu atributuetan. Kontuan hartu behar dugu `izena+aditza` konbinazio askotan euskarri-aditz jakin batzuk maiz agertzen direla: *egin, eman, hartu*....

VII.26 taulan, itxarondako kategoria-banaketatik gehien urruntzen diren 12 aditzak eman ditugu. Beraz, bada aditzaren eta UF-kategoriaren arteko korrelazio-maila bat. Interesgarria izan daiteke ikustea sailkatzaile automatikoak ezagutza horri etekina atera diezaiokeen. Aditza *string* motako atributua da, eta Wekaren `StringToVector` iragazkiaren bidez, zenbakizko atributu bihurtzen da.

Sailkatzaileak eratzeko eta trebatzeko ohiko prozedura izaten da ebaluazio-erreferentzia ikaste-multzo batean (*training test*) eta test-multzo (*test set*) batean ausaz bereiztea. Gure kasuan, kontuan hartu behar dugu ebaluazio-erreferentzia ez dela oso handia (1 145 instantzia edo adibide), eta, ondorioz, erreferentzia ausaz bitan bereiziz era daitezkeen ikaste/test multzo-bikoteen emaitzak uniformeak ez izateko arriskua dagoela. Zenbait proba egin ditugu hori egiaztatzeko, eta baieztatu dugu emaitzak ez direla egonkorrak, ausazko bereizketarekiko sentikorrak baizik. Beraz, komenigarria da ausazko zenbait bereizketa egitea, eta emaitzen batezbestekoa kalkulatzea. Hori da, hain zuzen ere, balidazio gurutzatuaren (*cross-validation*) prozeduran egiten dena.

aditza	id	col	free
egin	1	41	23
eman	2	20	3
eraiki	3	0	1
estutu	2	0	0
harrotu	2	0	0
ukitu	2	0	0
hartu	2	10	1
berotu	2	0	1
harrapatu	2	0	1
jarri	1	9	3
sartu	2	4	0
jaso	0	7	2

VII.26 Taula: UF kategorien araberako itxarondako maiztasun-banaketatik gehien urruntzen diren aditzak (Pearson-en $\chi 2$ testaren goranzko balioen araberako rankingeko lehen 12 aditzak. Hipotesi nuluaren araberako probabilitate-banaketa: `id` 0,070 / `col` 0,234 / `free` 0,696).

Balidazio gurutzatuan, erreferentzia n ataletan (*fold*) banatzen da, ausaz, eta n ebaluazio-txanda egiten dira. Horietako bakoitzean, $n-1$ atal erabiltzen dira ikasteko, eta gainerakoa da ebaluatzen den test-multzoa. Ebaluazioa m aldiz egin daiteke, ausaz sortutako n atal horiek m aldiz sortuz eta ebaluatuz, emaitzak are fidagarriagoak izan daitezen. Azkenean, txanda bakoitzean lortutako ebaluazio-neurrien batezbestekoa egiten da.

Sistema hori erabiltzen denean, ordea, kontuz ibili behar dugu atributuak automatikoki hautatzeko iragazkiekin. Izan ere, oso garrantzitsua da iragazkiak ikaste-multzoan dauden instantziak soilik aztertzea, test-multzotik independenteki. Bestela, ebaluatzeko ditugun datuak ari gara ikasteko erabiltzen. Balidazio gurutzatua erabiltzen denean, hainbat aldiz (ohikoena, 5 edo 10) egiten da ikaste/test multzo-bereizketa, eta, beraz, atributu-hautaketa horietako bakoitzean aplikatu behar da, uneko ikaste-multzoa kontuan hartuta. Atributu-hautaketa balidazio gurutzatua egin aurretik ezarriz gero, lehen aipatu dugun akatsa egingo dugu.

Wekak atributu-iragazketa automatikoa azaldu berri ditugun kondizioetan ondo egiteko eskaintzen dituen metodoen artean, *AttributeSelectedClas-*

sifier metasailkatzailea hautatu dugu[17].

Ebaluazio-laginean dagoen kategoria-banaketa (`id, col, free`) ez da uniformea (80 / 268 / 797). Adibide gehienak `free` kategoriakoak dira; beraz, sistemak kategoria horretakoak sailkatzen ikasiko du hobekien, eta gure interesa da, hain zuzen ere, `id` eta `col` sailkapena optimizatzea. Horregatik, saio batzuk egin ditugu *sampling* (laginketa) bidez `id` eta `col` kategoriakoen proportzioa handitzeko. Esaterako, adibide guztiak dituen datu-multzoaz gain, `free` kategoriakoen 1/3 eta 2/3 ausaz kendua duten datu-multzoak ere eratu ditugu. Esan gabe doa, laginketak ikaste-multzoari eragiten dio, ez test-multzoari, horrek datu errealak eduki behar baititu, ez sailkatzaileak hobeto ikasteko asmoz lagindu ditugunak. Baina balidazio gurutzatua erabiltzen denean, txanda bakoitzean ikaste-multzo eta test-multzo bikote desberdinak sortzen dira. Laginketa egitea ez da oso bideragarria kondizio horietan, eta alde batera utzi behar izan dugu.

Oinarri-lerrotzat erabili behar genukeen sailkatzailea dela eta, bi aukera hauek dira ohikoenak (Keller, 2003; Völker et al., 2007):

- Ausazko oinarri-lerroa (*random baseline, chance baseline*): kategoria ausaz esleitzen da.
- Gehiengoaren oinarri-lerroa (*majority baseline, frequency baseline*): ikaste-multzoan nagusi den kategoria esleitzen zaie test-multzoko instantzia guztiei.

Kategorien proportzioak desberdinak direnean, ausazko oinarri-lerroa ez da gomendagarria; horregatik, Kim eta Baldwinen (2007) antzera, gehiengoaren oinarri-lerroa erabili dugu. Bestetik, gehiengoaren oinarri-lerroa gainditzea zaila izan daiteke laginaren kategoria-banaketa orekatua izatetik aski urrun dagoenean. Hein batean, horrelakoxea da gure kasua, konbinazio libreen proportzioa beste biena baino nabarmen handiagoa baita.

Esperimentuetan, datu-multzo hauek erabili ditugu, jarraian azaltzen ditugun faktoreak hausnartu ondoren:

- `AM, DSim, MSFlex` eta `LFlex` datu-multzoak: esperimentu bakunetan, antzekotasun distribuzionaleko neurriek lortu dituzte emaitza onenak. Beraz, espero izatekoa da neurri horiekin osatutako datu-multzoa (`DSim`) darabilten sailkatzaileak hobeak izango direla idiomatikotasunaren propietate bakoitzeko datu-multzoak (`AM, MSFlex` eta `LFlex`) darabiltzaten sailkatzaileak baino; beraz, komeni zaigu hori egiaztatzea.

[17] http://weka.wikispaces.com/Performing+attribute+selection

- `4. osag.` datu-multzoa: interesgarria da aztertzea ea aurreko sailkatzaileen, bereziki DSim ezagutza darabiltenen, emaitzak hobetu daitezkeen lau propietateen neurketen datu-multzoa erabiliz. `AM, DSim, MSFlex` eta `LFlex` esperimentu bakunen emaitzak dituen datu-multzoa da.

- `4. osag.+ad.` datu-multzoa: aditzaren ekarria ere kontuan hartzea komeni da, sailkatzaileak aurkeztu berri dugun korrelazioaz balia daitezkeen egiaztatzeko. Konbinazioaren aditza ere atribututzat duen datu-multzoa da.

- Atributu-hautaketa: komeni da jakitea nolako emaitzak lortzen diren esperimentu bakunetan τ_B eta AP modalitate bakoitzean ($AP_{\texttt{UF}}$, $AP_{\texttt{id}}$, $AP_{\texttt{col}}$) emaitza onenak lortu dituzten neurriak erabiliz, atribututzat propietate bakoitzaren neurri guztiak erabili beharrean. Hautatze-lan hori egiteko, prozedura hauek erabili ditugu:

  - Eskuzko hautaketa
    - `eskuz_1` datu-multzoa: lau ezagutza-motetatik, τ_B eta AP modalitate bakoitzeko neurri onena aukeratu.
    - `eskuz_2` datu-multzoa: lau metriketako bakoitzeko onenetatik kopuru bat hartu (esaterako, 20). Neurri bakoitzari lau rankingetako posizioen araberako pisu bat eman, eta, pisuen baturen rankinga osatuz, 25 onenak aukeratu.
  - Hautaketa automatikoa
    - `CS-BF` datu-multzoa: Wekak atributuak automatikoki hautatzeko dituen iragazte-algoritmoetatik, osagai hauek dituen iragazkia erabili dugu *AttributeSelectedClassifier* metasailkatzailean: CfsSubsetEval1 ebaluatzailetzat[18], eta BestFirst bilaketa-metodotzat. Guztira, 37 atributu hautatu ditu iragazki horrek: 3 AM, 16 DSim, 7 MSFlex, 2 LFlex and 9 aditz.

VII.2.2 Emaitzak

VII.27 taulan ageri dira deskribatu berri ditugun datu-multzoekin lortutako emaitzak. Naive Bayes, j48 eta PART algoritmoen emaitzak ez ditugu bistaratu, beste hirurenak baino apalagoak direlako ia kasu guztietan.

Emaitza horietatik, honako alderdi hauek dira nabarmentzekoak:

[18] http://wiki.pentaho.com/display/DATAMINING/CfsSubsetEval

Atrib.	Metod.	CCI	$F_{\texttt{id}}$	$F_{\texttt{col}}$	$F_{\texttt{free}}$	F_{mikro}	F_{makro}
Oinarri-lerroa		69,607	0,000	0,000	0,821	0,571	0,274
AM	LR	70,218	0,024	0,164	0,823	0,613	0,337
	SMO	69,782	0,000	0,015	0,822	0,575	0,279
	RF	62,969	0,090	0,344	0,762	0,617	0,399
DSim	LR	74,061	0,270	0,468	0,842	0,714	0,527
	SMO	69,607	0,046	0,063	0,820	0,589	0,310
	RF	71,441	0,279	0,438	0,822	0,694	0,513
MSFlex	LR	71,004	0,236	0,326	0,822	0,665	0,461
	SMO	69,607	0,000	0,000	0,821	0,571	0,274
	RF	70,742	0,127	0,402	0,82	0,674	0,450
LFlex	LR	68,646	0,023	0,081	0,813	0,585	0,306
	SMO	69,694	0,000	0,007	0,821	0,573	0,276
	RF	69,432	0,000	0,157	0,816	0,604	0,324
4 osag.	LR	73,362	0,355	0,487	0,837	0,722	0,560
	SMO	76,070	0,300	0,479	0,858	0,731	0,546
	RF	73,712	0,336	0,475	0,840	0,719	0,550
4 osag.+ad.	LR	62,795	0,274	0,490	0,751	0,656	0,505
	SMO	**76,856**	**0,418**	**0,544**	**0,858**	**0,754**	**0,607**
	RF	73,974	0,304	0,447	0,843	0,713	0,531
eskuz_1	LR	72,751	0,265	0,361	0,842	0,688	0,489
	SMO	69,782	0,077	0,000	0,822	0,578	0,300
	RF	69,170	0,305	0,350	0,816	0,670	0,490
eskuz_2	LR	74,585	**0,353**	0,409	0,849	0,711	0,537
	SMO	69,869	0,000	0,022	0,824	0,579	0,282
	RF	70,917	0,343	0,404	0,822	0,691	0,523
CS-BF	LR	**75,721**	0,339	**0,487**	**0,854**	**0,732**	**0,560**
	SMO	73,450	0,149	0,390	0,838	0,685	0,459
	RF	72,838	0,364	0,435	0,836	0,709	0,545

VII.27 Taula: Ikasketa automatikoko esperimentuen emaitzak (LR: Logistic Regression; RF: Random Forest).

- Ezagutza-iturri bakarreko lau datu-multzoetatik, DSim da emaitza onenak dituena, sailkatzaile guztietan; zehazki, Logistic Regression da metodorik eraginkorrena DSim datu-multzoarekin. SMO metodoak oinarri-

lerroaren mailako emaitzak ditu `MSFlex` eta `LFlex` datu-multzoekin, eta, gainerakoekin, gutxigatik gainditzen du.

Ezagutza-iturri bakoitzaren eragina beste era batera ikustearren, taulan bistaratu ez ditugun esperimentu osagarri batzuk egin ditugu: `4 osag. + ad.` datu-multzotik iturri bakoitzeko atributuak ezabatu, eta emaitzetan gertatzen den inpaktua neurtu dugu. Ondorio nabariena da LFlex neurketak kentzeak dakarrela F neurriaren hobekuntza handiena. Esperimentu bakunetan lortutako emaitza eskasenak propietate honetakoak direla kontuan izanik (ikus VII.1.4.3 atala), koherentea da hori gertatzea. Interesgarriagoa da jakitea MSFlex neurriak ezabatuz gero emaitzetan eragin negatiboagoa gertatzen dela AM atributuak kenduz baino. Horrek MSFlex propietateak ikasteari AM propietateak baino gehiago laguntzen diola iradokitzen du, eta atributu-hautaketan gerta daitekeenaren zantzu batzuk ematen dituela uste dugu.

- Oro har, algoritmo guztiek emaitza hobeak dituzte idiomatikotasunaren ezagutza-iturriak konbinatuz (`4 osag.` eta `4 osag.+ad.` datu-multzoak), iturri bakarra erabiliz baino (salbuespen bakarra Logistic Regresionekin gertatzen da, `4 osag.+ad.` datu-multzoarekin: bistan dena, Logistic Regression erabiliz aditza atributuetan sartzeak zarata sortzen du). Horrek erakusten du idiomatikotasunaren izaera konplexua dela, eta aukeratutako lau propietateek bere ekarpena egiten dutela, neurri berean ez bada ere.

- Emaitza onenak SVM familiako SMO algoritmoarekin lortu dira, `4 osag.` eta `4 osag.+ad.` datu-multzoak erabiliz. Jakina da SVM familiako algoritmoek hobeto kudeatzen dutela atributu ugariak erabiltzeak sortu ohi duen zarata. Aditza kontuan hartzeak eragin positiboa du esapide idiomatikoen eta kolokazioen sailkapenean, baina ez du eraginik konbinazio libreenean.

- Ohartarazteko beste puntu bat da Logistic Regressionen emaitzak besteak baino ezegonkorragoak direla. SVM eta Random Forest erabilita, konbergentzia ez da erabilitako aldagai zaratatsuen kopuruarekiko dependentea (Biau, 2012). Beraz, atributuen hautaketa egoki batek hobetu litzake Logistic Regressionen emaitzak.

- Atributu hautatuen datu-multzoak direla eta, datuek argi erakusten dute CS-BF iragazkia darabilen atributu-hautaketa automatikoa hobeto dabilela eskuzko atributu-hautaketak baino. Automatikoki hautatuko atributu gehienak DSim propietatekoak dira (16), baina inte-

resgarria da MSFlex eta aditz-motako atributuek ere laguntzen dutela emaitzak hobetzen (hurrenez hurren, 7 eta 9); ekarpen txikiena AM atributuek (3) eta LFlex motakoek egite dute (2).

- Atributu-hautaketatik onura handiena ateratzen duen algoritmoa Logistic Regression da. Hain zuzen ere, CS-BF esperimentuetako F neurriaren mikrobatezbestekoa eta makrobatezbestekoa (0,732 / 0,560) dira gehien hurbiltzen direnak SMO metodoak `4 osag.+ad.` datumultzoarekin lortutako emaitza onenetara (0,754 / 0,607).

VII.3 Predikatu konplexu batzuk birsailkatzearen eragina

VI.4 atalean aurreratu genuenez, ebaluazio-erreferentzian badira konbinazio batzuk (esaterako, *jaramon egin, zor izan*), esapide erdiidiomatikotzat gabe, esapide idiomatikotzat jotzekoak liratekeenak. Arazo horren nondik-norakoak II.4 atalean eman genituen, euskarazko `izena+aditza` osaerako UFez jardutean. 17 konbinazio dira, eta uste dugu merezi duela horiek esapide idiomatikotzat sailkatzeak emaitzetan dakarren inpaktua zenbaterainokoa den egiaztatzea, batez ere konbinazio horien ezaugarrien irakasbidea izan daitekeelako. Horiek `id` gisa birsailkatuta, honelakoa da ebaluazio-erreferentziaren osaera: `id` 97; `col` 251; `free` 797.

VII.3.1 Esperimentu bakunak

II.4 atalean adierazi bezala, maiztasun handiko UFak dira, semantikoki erdikonposizionalak (izenak bere oinarrizko esanahia atxikitzen du), baina esanahi-unitate argiak direnak eta idiosinkrasia morfosintaktikoa dutenak (hedapenik gabeko aldakuntzetan, izena mugatzailerik gabe erabiltzen da: *jaramon egin, zor izan*).

Horrenbestez, aurresana litzateke AMen familiako neurri batzuek (f-k eta horrekin korrelazioa handia dutenek, hala nola t neurriak) eta MSFlex neurriek, birsailkatuak bi propietate horietan aski idiosinkratikoak direnez gero, emaitza hobeak izango dituztela $AP_{\texttt{col}}$ neurrian izan ezik. `id` kategoriako unitateen proportzioa handia ez denez, seguru aski Kendall τ_B-n aldaketa ez da asko nabarituko, baina bai $AP_{\texttt{id}}$-ren balioetan ($AP_{\texttt{UF}}$-renak ez dira aldatuko, $\texttt{id} \cup \texttt{col}$ bildura bera baita bi sailkapenetan).

VII.28 taulan, bi sailkapenen araberako Kendall τ_B koefizientearen balioak eta $AP_{\texttt{id}}$ eta $AP_{\texttt{col}}$-en balioak eman ditugu (idiomatikotasunaren osagai bakoitzaren neurketa onena). (1) sailkapena: aditu-taldeak egindakoa;

	neurria	τ_B(1)	τ_B(2)	$AP_{\texttt{id}}$ (1)	$AP_{\texttt{id}}$ (2)	$AP_{\texttt{col}}$ (1)	$AP_{\texttt{col}}$ (2)
	ausaz	0,000	0,000	0,070	0,086	0,234	0,219
AM	t neurria	0,197	0,201	0,084	0,171	0,383	0,317
	χ^2	0,037	-0,040	0,119	0,126	0,206	0,194
	$\texttt{MI}^3$	0,103	0,106	0,107	0,187	0,291	0,240
DSim	L2_Indri_hit_erl_NV	**0,322**	**0,325**	0,294	0,327	0,354	0,315
	L2_KL_hit_erl_NV	0,308	0,311	**0,320**	**0,334**	0,332	0,300
	L2_Indri_rankV_hazt	0,314	0,315	0,128	0,159	**0,431**	**0,402**
MSFlex	CPMI_osag_mugat	0,154	0,160	0,113	0,189	0,329	0,268
	CPMI_orok_ADJ	0,045	0,050	0,202	0,284	0,237	0,194
	CPMI_izena_mugat	0,133	0,139	0,101	0,168	0,331	0,278
	Hrel_ADJ	-0,055	-0,059	0,044	0,054	0,298	0,294
LFlex	z_PMI_V_WNhedap	0,110	0,108	0,120	0,128	0,259	0,247
	R_{nv}_EBantzdistr_20	0,066	0,072	0,066	0,102	0,323	0,283
	z_PMI_V_ELHWNhedap	0,101	0,098	0,122	0,130	0,250	0,237

VII.28 Taula: Kendall τ_B koefizientea, $AP_{\texttt{id}}$ eta $AP_{\texttt{col}}$-en balioak bi sailkapenetarako (idiomatikotasunaren osagai bakoitzaren neurketa onenak). (1): aditu-taldeak egindako sailkapenaren arabera; (2): 17 predikatu konplexu, `col` gisa gabe, `id` gisa sailkatuta.

(2) sailkapena: aipatu 17 predikatu konplexuak, `col` gisa gabe, `id` gisa sailkatuta.

τ_B koefizientearen balioetan hobekuntza txiki bat baino ez da nabari emaitza oneneko neurri gehienetan, LFlex neurketetan izan ezik, zeinetan beheranzko aldaketa gertatu baita. Bere txikian, MSFlex neurria da onen artean gehien igo dena, eta gero AM neurria. Ñabardura xehea izan arren, bat dator gure aurresanekin, birsailkatu ditugun konbinazioen ezaugarri nabarien artean aipatuak baititugu idiosinkrasia estatistikoa eta morfosintaxia. DSim neurriak direla eta, behatutako igoera ia hautemanezina da, eta `id` kategoriako UFen proportzioa igotzearekin asoziatua egon daiteke. Joera hori AP-ren balioetan baieztatzen den argitu behar dugu jarraian.

AP-ren emaitzak direla eta, lehenik kontuan izan behar dugu, `id` kategorian UF gehiago sailkatu ditugunez, (2) esperimentuen AP-ren oinarri-lerroa goraxeago dagoela, eta beheraxeago, berriz, `col` kategoriarena; dena den, aldea txikia da (0,016). Bestetik, gogora ekartzea komeni da VI.5.1 atalean adierazitakoa: $AP_{\texttt{id}}$ eta $AP_{\texttt{col}}$-en balioak ezin direla aldi berean arbitrarioki handiak izan, kontrajarrita baitaude.

Hori esanda, aldaketa nabarienak AM batzuen (t neurriaren eta MI^3-ren) $AP_{\texttt{id}}$-ren igoerak dira (0,086raino), eta orobat, edo areago, MSFlex neurketenak (0,9rainokoak). Horien $AP_{\texttt{col}}$-en balioei, jakina, jaitsiera bat dagokie, zertxobait txikiagoa. DSim neurketen $AP_{\texttt{id}}$-ren balioak, berriz, oso gutxi igotzen dira, 0,15-0,3 bitartean (oinarri-lerroaren igoeraren ondorio izan daitekeena). Azkenik, LFlexen kasuan, ez da joera uniformerik.

Horrenbestez, emaitza horiek argiago erakusten dute τ_B koefizientearen balioetan nabaritu dugun joera, eta berresten, beraz, birsailkatutako UFen ezaugarri ezagunak kontuan hartuz lehen egin ditugun aurresanak. Izan ere:

- f-rekin korrelazio handia duten AMen $AP_{\texttt{id}}$-ren balioak hainbeste igotzeak adierazten du birsailkatutako UFak estatistikoki idiosinkratikoak direla, eta, zehazki, maiztasun handikoak. Alde horretatik, horien profilak antz handiagoa du kolokazioen profilarekin esapide idiomatikoenarekin baino.

- DSim neurketetan dagoen aldeak ez dira asko urruntzen $AP_{\texttt{id}}$ eta $AP_{\texttt{col}}$-en oinarri-lerroen aldaketetatik; gainera, joera orokorra da aditzaren semantikarekiko neurketek izatea $AP_{\texttt{id}}$-ren balioen igoera handienak. Horrek adierazten digu birsailkatutako predikatuek aditzarekiko konposizionaltasun apalagoa dutela izenarekiko baino, hau da, erdikonposizional izatearen zantzuak dituztela; kolokazioei dagokien ezaugarria, hain zuzen ere.

- MSFlex neurketetan nabari da gehien $AP_{\texttt{id}}$-ren balioen igoera, birsailkatutako UFen propietate bereizgarrienetako bat idiosinkrasia morfosintaktikoa den seinale. Bi sailkapenekin egindako esperimentuetan, mugatasun-aldakuntzekiko eta `ADJ` hedapenarekiko malgutasunak dira parametro idiomatikoenak; hainbat parametrotan gertatu dira horien adinako igoerak, edo handiagoak, batik bat erlatiboaren kasuan (0,14rainokoak). Horrek guztiak iradokitzen du propietate honetan esapide idiomatikoen profila jo dugula.

Laburbilduz, esperimentalki baieztatu dugu 17 predikatu konplexu horien profila berezia dela: idiosinkrasia estatistikoari eta semantikoari dagokionez, erdiidiomatikoak dira, baina haien idiosinkrasia morfosintaktikoak esapide idiomatikoen kategorian kokatzen ditu. Nolako eragina du horrek guztiak sailkapen-atazan? Hori da jarraian aztertuko duguna.

VII.3.2 Ikasketa automatikoa

VII.2.2 atalean aurkeztu ditugun ikasketa automatikoko esperimentuak berregin ditugu sailkapen berriarekin. VII.29 taulan bistaratu ditugu sailkapen berriarekin lortutako emaitzak (VII.27 taulan emaitza onenak izandako datu-multzoekin egindako esperimentuenak bakarrik erakutsi ditugu). Parentesi artean, (2) sailkapenaren eta (1) sailkapenaren emaitzen kendura eman dugu. Kendura negatiboa denean, (1) sailkapenaren emaitzak hobeak dira.

Alderdi hauek dira nabarmentzekoak:

- (1) sailkapenarekin emaitza onenak lortu dituzten metodoak dira bigarrenean ere nagusitu direnak: SMO metodoa da onena, ezagutza-iturritzat `4 osag.` eta `4 osag.`+`ad.` datu-multzoak erabiliz; atributu-hautaketa eginez, berriz, CS-BF esperimentuetako Logistic Regressionen emaitzak dira onenak.

- Sailkatzaile eta datu-multzo horiekin lortutako emaitzetan, esan liteke, oro har, $F_{\texttt{id}}$ dela (2) sailkapenarekin hobetzen den metrika bakarra, (eta alderantzizkoa $F_{\texttt{col}}$-en kasuan). Hori espero izateko emaitza zen, (2) sailkapenean 17 instantzia gehiago baitira `id` kategoriakoak (beste hainbeste gutxiago `col` kategorian). Beraz, sailkatzaileak `id` kategoriako instantzia gehiago ditu ikasteko. Interesatzen zaiguna da horrek sailkapenaren emaitza globalean duen eragina, eta alde horretatik esan beharra dago gainerako metriketan (1) sailkapenaren emaitzak, oro har, hobeak direla.

Atrib.	Metod.	CCI		$F_{\texttt{id}}$		$F_{\texttt{col}}$		$F_{\texttt{free}}$		F_{mikro}		F_{makro}	
DSim	LR	73.537	(-0.524)	0.345	(0.075)	0.398	(-0.070)	0.846	(0.004)	0.705	(-0.009)	0.53	(0.003)
	SMO	69,782	(0,175)	0,076	(0,030)	0,008	(-0,055)	0,822	(0,002)	0,580	(-0,009)	0,302	(-0,008)
	RF	72,576	(1,135)	0,339	(0,060)	0,438	(-0,000)	0,837	(0,015)	0,708	(0,014)	0,538	(0,025)
4 osag.	LR	70,830	(-2,533)	0,365	(0,010)	0,443	(-0,044)	0,822	(-0,015)	0,700	(-0,022)	0,543	(-0,016)
	SMO	75,109	(-0,961)	0,375	(0,075)	0,418	(-0,061)	**0,855**	(-0,003)	0,725	(-0,006)	0,549	(0,004)
	RF	72,314	(-1,397)	0,382	(0,046)	0,395	(-0,080)	0,834	(-0,006)	0,703	(-0,016)	0,537	(-0,013)
4 osag.+ad.	LR	61,135	(-1,659)	0,286	(0,012)	0,409	(-0,081)	0,755	(0,004)	0,639	(-0,017)	0,483	(-0,022)
	SMO	**75,284**	(-1,572)	**0,428**	(0,010)	**0,485**	(-0,059)	0,853	(-0,005)	**0,740**	(-0,014)	**0,589**	(-0,018)
	RF	72,664	(-1,310)	0,301	(-0,003)	0,413	(-0,034)	0,836	(-0,007)	0,698	(-0,015)	0,517	(-0,015)
CS-BF	LR	**75,197**	(-0,524)	**0,387**	(0,048)	**0,45**	(-0,037)	**0,853**	(-0,001)	**0,725**	(-0,007)	**0,563**	(0,003)
	SMO	73,362	(-0,087)	0,222	(0,073)	0,378	(-0,012)	0,843	(0,005)	0,689	(0,004)	0,481	(0,022)
	RF	71,616	(-1,223)	0,307	(-0,057)	0,416	(-0,019)	0,83	(-0,006)	0,695	(-0,014)	0,518	(-0,027)

VII.29 Taula: (2) sailkapenarekin egindako ikasketa automatikoko esperimentuen emaitzak. Parentesi artean, (2) sailkapenaren eta (1) sailkapenaren emaitzen arteko kendura; negatiboa denean, (1) sailkapenaren emaitzak hobeak dira.

- Sailkatzaileek desberdin jasaten dute birsailkatzearen inpaktua. $F_{\texttt{id}}$ alde batera utzita, Logistic Regression da (1) sailkapenaren emaitzetatik (2) sailkapenaren emaitzetara jaitsiera handiena duena. Kontuan izanik metodo hori dela hiruetan zarata okerren kudeatzen duena, horrek iradokiko luke (2) sailkapenaren araberako datu-multzoa zaratatsuagoa dela.

Beste datu interesgarri bat da CS-BF iragazkiak hautatzen dituen atributuak kategoriaren arabera nola dauden banatuta (1) eta (2) sailkapenen kasuan. VII.30 taulan dugu informazio hori.

propietate-mota	**(1) sailk.**	%	**(2) sailk.**	%
AM	3	8,11	1	2,86
DSim	16	43,24	15	42,86
LFlex	2	5,41	2	5,71
MSFlex	7	18,92	9	25,71
ad	9	24,32	8	22,86
totala	37	100,00	35	100,00

VII.30 Taula: CS-BF iragazkiak hautatzen dituen atributu-kategorien kopuruak, (1) eta (2) sailkapenetan.

Bi banaketa horien arteko alde nabariena da (2) sailkapenarekin MSFlex atributu esanguratsuen kopurua handitu egiten dela, eta AM atributuena, berriz, gutxitu. Horrek adierazten du MSFlex propietateak, (2) sailkapenean, korrelazio handiagoa duela idiomatikotasunaren banaketarekin; hau da, 17 predikatuen malgutasun morfosintaktikoaren profila gertuago dagoela esapide idiomatikoen profiletik. AMen kasuan, berriz, (2) sailkapenean, AMek galtzen dute sailkapena doi egiteko esangura. Horrek adierazten du 17 predikatuen birsailkatzeak eragiten duela idiosinkrasia estatistikoa duten UFak sakabanatuago egotea `id` eta `col` kategorien artean (1) sailkapenean baino, eta AM neurriak ez direla hain eraginkorrak bi kategorien artean diskriminatzeko.

Horiek guztiak bat datoz esperimentu bakunek (1) eta (2) sailkapenekin izandako emaitzak komentatu ditugunean nabarmendu ditugun alderdiekin, eta baieztatu egiten dute, horrenbestez, birsailkatutako UFen profil berezia.

VIII. KAPITULUA

Ondorioak eta etorkizuneko lanak

Tesi-lan honetan, euskarazko `izena+aditza` egiturako unitate fraseologikoak (UFak) corpusetik automatikoki erauzi eta idiomatikotasun-mailaren arabera karakterizatzeko lan esperimentala egin dugu. Horretarako, marko teoriko bat definitu dugu, fraseologia konputazionalean gure egitekorako garatu diren teknikak aztertu ditugu, eta diseinu esperimental bat egin dugu. Ondoren, UF hautagaiak testutik automatikoki erauzteko eta forma kanonikoa esleitzeko prozesu automatikoa garatu dugu, eta 74 milioi hitzeko kazetaritza-corpus bat prozesatu. Gero, idiomatikotasunaren lau propietateak (idiosinkrasia estatistikoa, ez-konposizionaltasun semantikoa, finkapen morfosintaktikoa eta finkapen lexikala) zenbait metodoz kuantifikatzeko esperimentuak egin eta ebaluatu ditugu, *gold standard*tzat aditu-talde batek eskuz sailkatutako UF hautagai multzo bat erabiliz, zeinetan hiru kategoria bereizi baitira: esapide idiomatikoak, kolokazioak eta konbinazio libreak. Azkenik, xehe deskribatu ditugu emaitzak, aukeratutako bi atazetan (ranking-ataza eta sailkapen-ataza).

Kapitulu honetan, esperimentuen emaitzen analisitik ateratako ondorioak, egin ditugun ekarpenak eta etorkizunerako ikertze- eta garatze-aukerak azalduko ditugu.

VIII.1 Ondorio nagusiak

I.3 atalean, hiru ikergai hauek planteatu genituen:

(i) Idiomatikotasunaren eta bere propietate bakoitzaren neurketen artean dagoen korrelazioa aztertzea.

(ii) UFen propietateen ebidentzia enpirikoak zenbateraino datozen bat teoria fraseologikoak UFentzat oro har zein UF kategoria bakoitzarentzat aurresandakoarekin.

(iii) UFak automatikoki sailkatzeko, haien propietateen kuantifikazioaren emaitzak ikasketa automatikoko sistema batean konbinatzeak duen eraginkortasuna ebaluatzea.

Hiru ikergai horiei begira, hauek dira, esperimentuen emaitzak aztertu ondoren, atera ditugun ondorio nagusiak:

- Idiomatikotasunaren osagai diren propietate guztiek, hein desberdinean bada ere, ranking-atazako oinarri-lerroa gainditzen duen korrelazioa dute ebaluazio-erreferentziako hautagaien idiomatikotasun-mailarekin. Halaber, sailkapen-atazan, beren ekarpena egiten dute, propietate desberdinetako atributuez osatutako datu-multzoen emaitzek argi gainditzen baitituzte propietate bakoitzeko datu-multzoenak. Horrek guztiak **idiomatikotasunaren izaera konplexua** salatzen du.

- Idiomatikotasun-mailarekin ranking-atazako τ_B korrelazio onena duten neurriak konposizionaltasun semantikoa neurtzen duten antzekotasun distribuzionaleko neurriak dira (DSim - *distributional similarity*). Zehazki, IR arloko Indri indizeak eta Kullback-Leibler dibergentziak dituzte emaitza onenak. Sailkapen-atazan ere, propietate bereko datu-multzoekin egindako esperimentuetan, `DSim` datu-multzoarekin lortzen dituzte sailkatzaileek bere emaitza onenak. **Idiosinkrasia semantikoaren nabarmentasuna** erakutsi du horrek, eta inplikazio hauek ditu:

  - Gure marko teorikoaren aurresanetako bat baieztatu du horrek: konposizionaltasun-maila gradualki gutxituz doa esapide idiomatikoetatik kolokazioetan barrena konbinazio libreetaraino doan continuumean; hau da, konbinazioa konposizionala ez izatea ez da esapide idiomatikoen ezaugarri esklusiboa.
  - Ranking-atazan DSim neurri onenak IR arloko bi indize izatea, arlo honetan erabili ohi diren beste teknika batzuen gainetik, nabarmentzeko emaitza da, eta gure ikerketaren ekarpentzat dugu.
  - DSim esperimentuek UFen erauzketan ohikoen eta estandarren den elkartze-neurrien (AM - *association measures*) bidezko neurketaren emaitzak gainditzen dituzte. Etorkizunean, ataza honetan oinarritutako aplikazio bat garatzerakoan, emaitza hau kontuan hartu egin behar litzateke.

- Ranking-atazan batez besteko doitasunaren (*AP* - *average precision*) bidez neurtu dugun UF-kategoria bakoitzarekiko portaera dela eta, emaitzek agerian jarri dituzte esapide idiomatikoen eta kolokazioen ezaugarrien inguruko alderdi esanguratsu batzuk:

  – Esapide idiomatikoen ranking onenak DSim esperimentuetan lortu ditugu, bigramaren semantika izenaren eta aditzaren semantikekiko konparazioa egiten duten neurriak erabiliz. Baina kolokazioen kasuan, horrelako neurriak AM onenen eta malgutasun morfosintaktikoaren neurketa (MSFlex) onenen arteko emaitzak dituzte, hau da, ez dira nagusitzen. Bai, ordea, bigramaren semantika aditzaren semantikarekin konparatzen denean. Horrek baieztatzen du teoriak aurresandako **kolokazioaren erdikonposizionaltasuna**, eta horren gakoa aditzaren semantika berezian dagoela.

  – Esapide idiomatikoen eta kolokazioen ranking-emaitzetan kontraste handiena erakutsi duen propietatea malgutasun morfosintaktikoa da: bigarren AP_{id} onena du, baina AP_{col} apalenetakoa. Horrek **kolokazioen malgutasun morfosintaktiko handia** jarri du agerian, eta propietate hori dutela kolokazioek bereizgarritasun txikienekoa; marko teorikoan aurresandakoarekin bat dator ondorio hori.

- Erakutsi dugu zenbait predikatu konplexuren (*jaramon egin, zor izan* modukoen) profila berezia dela: esapide erdiidiomatikoen (kolokazioen) moduko idiosinkrasia estatistikoa eta erdikonposizionaltasuna dute, baina esapide idiomatikoen moduko finkapen morfosintaktikoa. Horren agergarri enpirikoa da AM eta MSFlex bidezko neurketak sentikorrak direla predikatu konplexu horien sailkapenean erabili diren irizpide desberdinekiko, eta emaitzetan deskribatu ditugun aldeak profil berezi horren aldeko ebidentziak dira. Horrek esan nahi du horrelakoak ez direla egoki kokatzen erabili dugun idiomatikotasunaren continuumean. Horiek horrela, kontuan hartzeko aukera bat izan liteke horrelakoei berariazko kategoria esleitzea, esapide idiomatikoen eta kolokazioen artekoa izan litekeena.

- Konbinazioaren aditza kontuan hartzeak sailkapen automatikoa hobetzen duela erakutsi dugu, batez ere esapide idiomatikoen eta kolokazioen kategoriak hobeto sailkatzen ikasteko. Ebaluazio-erreferentzia aztertuz, ikusi dugu badela korrelazio bat aditz batzuen eta dagokion

konbinazioa UF izatearen artean. Korrelazio handienekoen artean, ohiko euskarri-aditzak dira ugarienak (*egin, eman, hartu, jarri, sartu...*). Hain zuzen ere, ikasketa automatikoan atributuak automatikoki aukeratzeko erabili dugun iragazkiak ere multzo horretako aditzak hautatu ditu batik bat.

- Sailkapen automatikoan, emaitza onenak SVM familiako SMO algoritmoarekin lortu dira, idiomatikotasunaren ezagutza-iturri guztiak eta aditzaren informazioa konbinatuz (`4 osag.+ad.` datu-multzoa). Kontuan hartzen badugu datu-multzo horiek osatzeko egin behar den datu-prozesatzea eta kalkulua aski handiak direla, atributu-hautaketa automatikoko esperimentuetan Logistic Regression algoritmoarekin lortutako emaitza erlatiboki onek atea irekitzen diote, aplikazio praktikoari begira, bideragarritasunaren eta emaitzen kalitatearen arteko erlazio on bat lortzeko aukerari.

- Malgutasun morfosintaktikoko (MSFlex) eta, batez ere, malgutasun lexikaleko (LFlex) esperimentuen emaitzak ez dira espero zitezkeen mailakoak, kontuan hartuta fraseologia teorikoak UFen ezaugarri nabarmenetakotzat jotzen dituela, eta III.9 zein III.10 ataletan aurkeztutako lan esperimental batzuetan (Fazly eta Stevenson, 2007; Van de Cruys eta Moirón, 2007) adierazi dela malgutasun sintaktikoak eta lexikalak UFak karakterizatzeko beste teknikek baino emaitza hobeak dituztela. MSFLex neurketen kasuan bederen, emaitza aipagarriak lortu dira esapide idiomatikoen ranking-atazan, eta, sailkapen automatikoan, bigarren atributu-multzo hautatuena izan da, DSim neurketen atzetik. Baina LFlex neurketak beste propietateen azpitik agertzen dira metrika guztietan. Beraz, gure esperimentuek ez dute aipatu egileen mailako baieztapenik egiteko emaitzarik izan.

  Hori esplikatzeko, hipotesi batzuk plantea daitezke:

  – MSFlex esperimentuen kasuan, euskarazko UFak morfosintaktikoki malguagoak izan litezke beste hizkuntzetakoak (ingelesa eta nederlandera) baino, edo, antzeko efektua izan lezakeen alderantzizko hipotesia eginez, euskarazko konbinazio libre batzuk ez lirateke espero bezain malguak. Bigarren horren arabera, esan liteke malgutasun morfosintaktiko mugatua izatea badela UFen ezaugarri bat, baina ez ezaugarri diskriminatzaile edo bereizgarria, konbinazio libre batzuk ere halatsukoak baitira. Esaterako, mugatasun-aldakuntzak direla eta, izena batez ere modu gene-

rikoan erabiltzen den *ogia erosi* edo *egunkaria irakurri* moduko konbinazio libre batzuetan, badirudi izena pluralean ohiko konbinazioetan baino gutxiago erabiltzen dela. Baina konbinazioak ez dira horregatik idiomatikoagoak.

- Euskara, batez ere kazetaritza-erregistroan, gutxiago finkatua izan liteke, eta, beraz, kolokatiboaren aukera lexikala ez litzateke hain argi finkatua. Aditu batzuek adierazi dutenez (Altzibar, 2005: 12-13), kazetaritza-jardunean kolokazio berri asko sortzen ari dira; horren ondorio bat izan daiteke horietako batzuk sinonimoak izatea (*dimisioa aurkeztu/eman*; *gola egin/sartu*; *elkartasuna adierazi/agertu/azaldu/erakutsi/eman*); edo, bestela, lehendik tradizioan dauden batzuen lehiakideak izatea (*atentzioa eman/deitu*, *bilera egin/ospatu*). Litekeena da tradizio handiagoko edo erregistro egonkorragoko testuez osatutako corpus bat erabiliz aukera lexikal finkoagoko kolokazioak erauztea.
- Ordezkagarritasuna egiaztatzeko erabili diren euskarazko baliabideak kalitate eta estaldura gutxikoak izatea. Van de Cruys eta Moirónek (2007) ohartarazten dute emaitzak sentikorrak direla erabiltzen diren ordezkoen multzoen kalitatearekiko. Ikusi dugu erabili ditugun baliabideen estaldura mugatuaren ondorioz UF hautagai askotxoren malgutasuna ezin dela neurtu, ordezkorik edo ordezkoen konbinazioen corpus-agerpenik ezaren ondorioz; gainera, konparaziotzat hartu ditugun ikerlanetan erabili diren baliabideak ez bezalakoak dira (izenaren eta aditzaren arteko erlazio sintaktikoaren informaziorik ez dugu). Huts egiteko beldur gabe esan dezakegu horrek denak eragina izan duela emaitzen kalitatean, eta etorkizuneko lan-ildoetarako ideia-iturri izan behar du nahitaez.
- Antzekotasun distribuzionala neurtzeko, Fazly eta Stevensonek (2007) kosinua darabilte, eta, gure esperimentuetan, neurri horren portaera beste batzuena baino nabarmenki okerragoa da. Beraz, gure emaitzen ikuspegitik, zentzuzkoa da haien emaitzetan kosinuaren bidezko DSim neurketa ez izatea onena, baina DSim neurtzeko erabilitako neurriari zor zaio beharbada hori.

Aitortu beharra dago aurreko hipotesiek baduketela espekulaziotik; beraz, egiaztapen esperimentala eta analisi sakonagoa eskatzen dute. Eginbehar horrek etorkizunerako ikertze-aukera batzuk zabaltzen ditu, gainera. VIII.3 atalean azalduko ditugu interesgarri eta bideragarrienak.

VIII.2 Ekarpenak

Hauek dira tesi-lan honen ekarpen nagusiak:

- Euskarazko UFen idiomatikotasunaren definizio operatibo bat eta sailkapen-eredu bat landu ditugu, eta eredu hori UFen erauzketa eta karakterizazio automatikoa egiteko markoan erabili da. Definizio- eta sailkapen-eredua maila teorikoan ez ezik, praktikoan ere gauzatu da.

- Corpus handi bat bildu eta prozesatu da (74 milioi hitz), eta erauzitako bigrama-bilduma eta dagokion informazio estatistikoa oso baliagarriak dira hiztegigintzan zein hizkuntza-teknologien arloko hainbat atazatan erabiltzeko. Aplikazio errealei begira, informazio hori balio handikoa da, etekin eta performantzia oneneko aukerak inplementatzeko.

- Ebaluaziorako bi erreferentzia landu dira, hiztegi-erreferentzia bata eta eskuz landutako *gold standard*a edo erreferentzia sailkatua bestea. Bi erreferentzia horiek baliagarriak dira ikerkuntza-proiektuaren hurrengo urratsetarako eta ikertze-ildo berrien etorkizuneko lanetarako; eta, horretatik kanpo ere, ebaluazio-erreferentziatzat erabil litezke.

- Ebaluazio-erreferentzia sailkatua eratzeko prozedura diseinatu eta argibideak zehaztu dira. Aditu sailkatzaileen arteko adostasuna (ITA – *inter-tagger agreement*) kalkulatu da, eta adostasun moderatua lortu da. Horrek egitekoaren zailtasuna jarri du agerian, eta etorkizunean irizpideen lanketan sakondu beharra.

- Testuetatik `izena+aditza` egiturako UFak automatikoki erauzteko eta aldakuntzak forma kanonikora normalizatzeko metodologia bat landu da, eta hori gauzatzeko oinarrizko tresnak garatu.

- UFen idiomatikotasunaren osagai diren propietateak (idiosinkrasia estatistikoa, ez-konposizionaltasun semantikoa, finkapen morfosintaktikoa eta finkapen lexikala) neurtzeko teknikak ikertu eta euskararako egokitu eta garatu ditugu.

- Propietate esanguratsuena den idiosinkrasia semantikoa (ez-konposizionaltasuna) neurtzeko, IR arloko teknikak erabili ditugu, eta horien bidez lortu ditugu emaitza onenak, gainera. Arlo horretako metodoak egiteko honetan erabili diren lehen aldia da.

- UF hautagaien sailkatzaile automatikoak garatzea eta ebaluatzea. Aplikazio praktiko bati begira, ekarpen handiena egiten duten propietateen

eta haiek neurtzeko metodoen informazioa eskuratu dugu. Ebidentzia horietan oinarrituta, testatu diren metodoetatik batzuk baztertu eta beste batzuk inplementatzeko erabakiak hartu ahal izango dira.

- Garatutako metodoa baliozkoa da, eta erraz egokitu daiteke beste egitura morfosintaktiko batzuetako bigramak erauzteko (`izena+izena`, `izena+izenondoa`, `izenlaguna+izena`...). Tesi-lan honetan garatutako teknika batzuk egokitu ditugu Elhuyar Fundazioaren *Web-corpusen Atari*ko euskarazko corpus elebakarra prozesatzeko (125 miloi hitz), eta, `izena+aditza` ez ezik, `izena+izenondoa` eta `izena+izena` bigramak ere forma kanonikoan erauzi eta AMen bidez estatistikoki prozesatu ditugu. Emaitzak doan kontsulta daitezke atariko "Hitz-konbinazioak" atalean[1]. Bigramen adibide bana bistaratzen da, baina erabiltzaileak denak kontsultatzeko aukera du. Horretarako, aski da konbinazioaren gainean, edo, bestela, eskuinean dagoen aukeran, klik egitea. C eranskinean, kontsulta-interfaze horretan lor daitekeen informazioaren erakusgarri bat eman dugu.

VIII.3 Etorkizuneko lanak

Euskarazko fraseologia konputazionaleko lanak hastapenetan daude eta, tesi-lan honetan ere, ikertze-ildoaren lehen urratsak egin ditugu. Beraz, ikergaian aurrera egiteko eta sakontzeko bide zabala eta luzea dago. Gainera, ikerkuntzaren ondorioetan adierazi dugunez, lortutako emaitza batzuetarako proposatu ditugun azalpenak egiaztatzeko aukera ere bilatu behar dugu etorkizuneko lanetan. Honako hauek dira aukera interesgarri eta bideragarrienak:

- Informazio sintaktiko aberatsagoa duen corpus etiketatu bat erabiltzea. Maltixa erabiliz (Bengoetxea eta Gojenola, 2010), perpausaren osagaien arteko mendekotasunen informazioa eskura genezake. Horrek bide emango luke hainbat alderdi ikertzeko:
  - Hautagaien erauzketa, agerkidetza hutsean gabe, erlazio sintaktikoan oinarritzea, Seretanek (2011) proposatzen duen bidetik. Erauzketa doiagoa egiteko aukera eman lezake horrek.
  - Malgutasun morfosintaktikoaren neurketan aldakuntzen erauzketaren estaldura eta doitasuna hobetzea, lan honetan garatutako

[1] http://webcorpusak.elhuyar.org/cgi-bin/kolokatuak.py

azaleko gramatikaren mugak gainditzuz, Bannardek (2007) egiten duen eran.

  - Malgutasun lexikalaren kasuan, Linen (1999) antzeko thesaurus bat eratzea, eta horko itemak erabiltzea ordezkagarritasuna neurtzeko. Horrek gutxitu lezake, beharbada, tesi-lan honetan behatu dugun fenomeno bat: malgutasun lexikalaren neurketak sentikorrak izatea osagaien ordezkoak lortzeko erabiltzen diren baliabide lexikalen kalitatearekiko.

- Beste corpus-mota batzuk erabiltzea: literarioak, espezializate-arlokoak, web-corpusak... Helburua da ondorioak sendoak diren egiaztatzea (berresten diren ala aldatzen diren). Adibidez, literatura oso interesgarria da LFlexen emaitza eskasak esplikatzeko proposatu ditugun ideia batzuk egiaztatzeko. Gainera, LFlexen kasuan, aukera emango luke lan honetan erabilitako corpus-motak emaitzetan izan duen eragina haztatzeko; hain zuzen ere, egiazta genezake espero bezain emaitza onak lortu ez izanaren esplikaziotzat planteatu dugun kazetaritza-erregistroaren normalizazio-maila konparatiboki apalaren hipotesia.

- Ebaluazio-erreferentzia handiagoa eratzea, eta adituen arteko adostasuna handitzea, idiomatikotasunaren izaeraren eta UFen sailkapenaren irizpideak areago landuz eta finduz. Horrek sailkatzaile hobeak eraikitzeko aukera emango luke.

- Beste egitura sintaktiko batzuetako UFak erauztea eta idiomatikotasunaren arabera karakterizatzea: `izena+izena`, `izena+izenondoa`, `izenlaguna+izena`, `adberbioa+aditza` eta `adberbioa+izenondoa`. Lan horren lehen urrats batzuk egin dira; esaterako, Elhuyar Web-corpusetik `izenondoa+izena` eta `izena+izena` konbinazioak erauzi dira, baina agerkidetza-teknikak baino ez dira inplementatu. Bestetik, `izena+aditza` konbinazioen kasuan, komeni da egiaztatzea konposizionaltasunaren eta malgutasun morfosintaktikoaren neurketek dakarten hobekuntza egitura horietan ere gertatzen den.

- Interpretazio literala eta idiomatikoa izan ditzaketen konbinazioen agerpenak bereiztea. Egiteko hau euskarazko UFei aplikatzea etorkizuneko lantzat du aipatua Urizarrek bere tesian (Urizar, 2012: 316); eta aplikatzea proposatzen duen teknika batzuek erlazio zuzena dute gure ikerkuntzan aplikatu eta garatu ditugunekin; zehazki, agerpenaren propietate semantikoak, eta malgutasun morfosintaktikoa zein lexikala neurtzekoak. Gakoa da, literal/idiomatiko bereizketan, konbinazioaren

agerpen bakoitza ebatzi behar dela, konbinazioaren portaera orokorra karakterizatu beharrean. Erabili ditugun teknikak ataza berrira egokitu egin behar dira, beraz[2]. Ataza konplexua da, agerpen bakoitzari buruzko informazioa murritza izaten baita.

- Corpus paraleloak erabiltzea, zenbait helburutarako:
  - Ordezkagarritasunaren azterketa konparatiboa egitea. Helburua litzateke egiaztatzea euskarazko kolokazioen malgutasun lexikala, batez beste, handiagoa den beste hizkuntza batzuetakoena baino (es, en, fr, de); zeren, hala izatera, horrek esplika bailezake euskarazko UFen malgutasun lexikalaren neurketan izan ditugun emaitzak espero baino apalagoak izatea.
  - UFen erauzketa elebiduna egitea, hau da, corpusetik bi hizkuntzatako UF bikote elebidunak erauztea eta karakterizatzea. Aplikazio zuzena duen lana da, oso erabilgarria baita emaitza hainbat arlotan: itzulpengintza automatikoan, ordenagailuz lagundutako itzulpen-aplikazioetan, hiztegigintzan, eta itzulpenaren ikerkuntzan. Orain arte, ElexBi termino-erauzle elebiduna da (Alegria et al., 2006) corpus paraleloen arlo honetako ustiaketan garatu dugun tresna bakarra; fraseologia elebiduna erauzteko aukera gehitzea da proposatzen den lanaren helburua.
- Euskararako lortu ditugun emaitzak beste hizkuntza batzuetan lortutako emaitzekin kontrastatzea, interes berezia jarriz euskararenak bezalako edo haren antzeko ezaugarri morfosintaktikoak dituzten hizkuntza eranskarietan, edo antzeko egoera soziolinguistikoan dauden hizkuntza minorizatuetan. Azterketa kontrastibo horiek zenbait alderdiren inguruko datu interesgarriak eman ditzakete, hala nola etiketatze sintaktikoak erauzketaren eta karakterizazioaren emaitzetan duten eragina, eta normalizazio-prozesuaren aurreratze-mailak kolokazioen malgutasun lexikalean duen eragina.
- Tesi-lan honen eta azaldu berri ditugun etorkizuneko hainbat lanen helburu nagusietako bat da UFak erauzi eta karakterizatzeko tresnak garatzea. Zenbait jarduera eta arlotan aplikazioa izango lukete tresna

[2]Esperimentu horietako batzuek egokitzapen gutxi behar lituzkete. Esaterako, konposizionaltasuna neurtzeko Lemurrekin egin ditugun L2 modalitateko esperimentuetan, konbinazioaren agerpenak banan-banan konparatu dira osagaien testuinguruez osatutako bildumarekin. Konparazio horien emaitzak konbinatu beharrean, emaitza bakoitza independenteki prozesatu behar litzateke, agerpena ebaluatzeko.

horiek, eta nabarmenena da hiztegigintzan, itzulpengintzan eta hizkuntzaren prozesamendu automatikoko hainbat atazatan behar diren baliabide lexikal aberatsak eratzea, elikatzea eta egunean edukitzea. Arlo horietan, esapide idiomatikoak eta kolokazioak eskuratzea eta bereiztea funtsezkoa da. Hiztegien kasuan, hiztegigintza automatizaturantz eraman gintzakeen bidea geltoki horietatik iragango litzateke, dudarik gabe.

Bibliografia

Aduriz, I., Agirre, E., Aldezabal, I., Alegria, I., Arregi, X., Arriola, J. M., Artola, X., Gojenola, K., Maritxalar, A., Sarasola, K., et al. (2000). A word-grammar based morphological analyzer for agglutinative languages. *Proceedings of the 18th Conference on Computational Linguistics* 1. lib., 1–7. Association for Computational Linguistics.

Agirre, E. eta Edmonds, P. (2006). *Word Sense Disambiguation*. Springer.

Agirre, E., Martínez, D., de Lacalle, O. L., eta Soroa, A. (2006). Two graph-based algorithms for state-of-the-art WSD. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 585–593. Association for Computational Linguistics.

Agirre, E. eta Soroa, A. (2007). Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. *Proceedings of the 4th International Workshop on Semantic Evaluations*, 7–12. Association for Computational Linguistics.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data* 656. lib. John Wiley & Sons.

Aisenstadt, E. (1981). Restricted collocations in English lexicology and lexicography. *ITL. Review of Applied Linguistics*, 53:53–61.

Aldezabal, I., Ansa, O., Arrieta, B., Artola, X., Ezeiza, A., Hernández, G., eta Lersundi, M. (2001). EDBL: A general lexical basis for the automatic processing of Basque. *IRCS Workshop on linguistic databases*, 1–10, Philadelphia, AEB.

Aldezabal, I., Aranzabe, M. J., de Ilarraza, A. D., Estarrona, A., Ezeiza, N., eta Uria, L. (2011). Corpusen etiketatze linguistikoa. *Anuario del Seminario de Filología Vasca Julio de Urquijo*, 43(1–2):37–50.

Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., eta Urizar, R. (2004a). Representation and treatment of multiword expressions in Basque. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, 48–55. Association for Computational Linguistics.

Alegria, I., Aranzabe, M., Ezeiza, A., Ezeiza, N., eta Urizar, R. (2002). Robustness and customisation in an analyser/lemmatiser for Basque. *Proceedings of Workshop on "Customizing knowledge in NLP applications". Third International Conference on Language Resources and Evaluation*, 1–6, Las Palmas de Gran Canaria.

Alegria, I., Gurrutxaga, A., Lizaso, P., Saralegi, X., Ugartetxea, S., eta Urizar, R. (2004b). A XML-based term extraction tool for Basque. *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*, 1733–1736.

Alegria, I., Gurrutxaga, A., Saralegi, X., eta Ugartetxea, S. (2006). Elexbi, a basic tool for bilingual term extraction from Spanish-Basque parallel corpora. *Proceedings of the 12th International Congress of Lexicography - EURALEX '06*, 159–165, Turin.

Allan, J., Callan, J., Collins-Thompson, K., Croft, B., Feng, F., Fisher, D., Lafferty, J., Larkey, L., Truong, T., Ogilvie, P., et al. (2003). The Lemur Toolkit for Language Modeling and Information Retrieval. *The Lemur Project*.

Alonso, M. (1996). Hacia una generalización de la información colocacional en un léxico formal. *Procesamiento del Lenguaje Natural*, 19:52.

Alonso, M. (1998). *Étude sémantico-syntaxique des constructions à verbe support*. Doktorego-tesia, Universite de Montreal.

Alonso, M. (2000). Verbos de apoyo, funciones léxicas y traducción automática. *Revista de Lexicografía*, 6:155–178.

Alonso, M. (2004). Elaboración del diccionario de colocaciones del español y sus aplicaciones. *De lexicografia: actes del I Symposium Internacional de Lexicografia* 15. lib., 149–162, Bartzelona. Documenta Universitaria.

Altuna, P. eta Casares, J. A. M. (2003). *Arnaud Oihenart: Euskal atsotitzak eta neurtitzak*. Euskaltzaindia.

Altzibar, X. (2005). Kolokazioak euskaraz. Zer axola duten kazetaritzan. *Euskarazko kazetaritzaren I. kongresua. Kazetaritza euskaraz: oraina eta geroa.*, 383–395. UPV/EHU.

Altzibar, X. (2012). Fraseologismoen hiztegi baten beharraz. Lakarra Andrinua, J. A., Gorrochategui Churruca, J., eta Urgell Lázaro, B.(ed.), *Koldo Mitxelena Katedraren II. Biltzarra*, 549–574. UPV/EHU.

Altzibar, X. (2013). Hizketa formula ohikoak. *Anuario del Seminario de Filología Vasca Julio de Urquijo*, 53–71.

Amosova, N.Ñ. (1963). *Osnovui anglijskoy frazeologii.* Leningrad: University Press.

Areta, N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., De Ilarraza, A. D., Ezeiza, N., eta Sologaistoa, A. (2007). ZT Corpus: Annotation and tools for Basque corpora. *Proceedings of the Corpus Linguistics Conference 2007.* University of Birmingham.

Arrieta, B. (2010). *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean.* Doktorego-tesia, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU, Donostia.

Azkarate, M. (1990). *Hitz elkartuak euskaraz.* Filosofi-Letren Fakultatea, Deustuko Unibertsitatea, Donostia.

Azkue, R. M. d. (1989). Euskalerriaren Yakintza. Literatura popular del País Vasco. *Euskaltzaindia-Espasa Calpe.*

Baayen, R. H. (2001). *Word frequency distributions* 18. lib. Kluwer Academic Pub.

Baeza-Yates, R. A. eta Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, AEB.

Baldwin, T. (2006). Compositionality and multiword expressions: Six of one, half a dozen of the other. *Invited talk given at the COLING/ACL06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties.* Association for Computational Linguistics.

Baldwin, T., Bannard, C., Tanaka, T., eta Widdows, D. (2003). An empirical model of multiword expression decomposability. *Proceedings of the ACL 2003 Workshop on Multiword Expressions: analysis, acquisition*

and treatment, 89–96, Sapporo, Japonia. Association for Computational Linguistics.

Baldwin, T. eta Kim, S. (2010). Multiword expressions. Indurkhya, N. eta Damerau, F. J.(ed.), *Handbook of Natural Language Processing, second edition*, 267–292. CRC Press, Taylor and Francis Group, Boca Raton, AEB.

Bally, C. (1909). *Traité de stylistique française* 1. lib. Librarie Klincksieck, Paris, 3. ed. [1951] edition.

Banerjee, S. eta Pedersen, T. (2003). The design, implementation, and use of the Ngram Statistics Package. *Computational Linguistics and Intelligent Text Processing*, 370–381. Springer.

Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, 1–8. Association for Computational Linguistics.

Bannard, C., Baldwin, T., eta Lascarides, A. (2003). A statistical approach to the semantics of verb-particles. *Proceedings of the ACL 2003 Workshop on Multiword Expressions: analysis, acquisition and treatment* 18. lib., 65–72. Association for Computational Linguistics.

Barkema, H. (1994a). Determining the syntactic flexibility of idioms. *Realising and Using English Language Corpora*, 39–52.

Barkema, H. (1994b). The idiomatic, syntactic and collocational characteristics of received NPs: some basic statistics. *Hermes*, 13:19–40.

Barkema, H. (1996). Idiomaticity and terminology: A multi-dimensional descriptive model. *Studia Linguistica*, 50(2):125–160.

Bauer, L. (1983). *English word-formation*. Cambridge University Press.

Béjoint, H. (2000). *Modern Lexicography: An introduction*. Oxford Universty Press.

Bengoetxea, K. eta Gojenola, K. (2010). Application of different techniques to dependency parsing of Basque. *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, 31–39. Association for Computational Linguistics.

Benson, M., Benson, E., eta Ilson, R. (1986). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. John Benjamins, Amsterdam & Philadelphia.

Berry, M. W., Dumais, S. T., eta O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595.

Berry-Rogghe, G. (1973). The computation of collocations and their relevance to lexical studies. *The Computer and Literary Studies*, 103–112.

Berry-Rogghe, G. (1974). Automatic identification of phrasal verbs. *Computers in the Humanities*, 16–26.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 98888:1063–1095.

Biemann, C. eta Giesbrecht, E. (2011). Distributional semantics and compositionality 2011: Shared task description and results. *Proceedings of the Workshop on Distributional Semantics and Compositionality*, 21–28. Association for Computational Linguistics.

Biemann, C. eta Nygaard, V. (2010). Crowdsourcing Wordnet. *Proceedings of the 5th Global WordNet Conference*, Mumbai, India. ACL Data and Code Repository, ADCR2010T005.

Bolboacă, S. eta Jäntschi, L. (2006). Pearson versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, 9:179–200.

Bolinger, D. (1977). Idioms have relations. *Forum linguisticum*, 2(2):157–169.

Bosque, I. (2001). Sobre el concepto de colocación y sus límites. *Lingüística Española Ectual*, 23(1):9–40.

Bouma, G. (2010). Collocation extraction beyond the independence assumption. *Proceedings of the ACL 2010 Conference Short Papers*, 109–114. Association for Computational Linguistics.

Breiman, L. (2001). Random Forests. *Machine learning*, 45(1):5–32.

Buckland, M. K. eta Gey, F. C. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1):12–19.

Burger, H. (1998). *Phraseologie: eine Einführung am Beispiel des Deutschen.* Berlin: Erich Schmidt.

Carreras, X., Chao, I., Padró, L., eta Padró, M. (2004). Freeling: An open-source suite of language analyzers. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04).*

Chernuisheva, I. (1964). *Die Phraseologie der gegenwärtigen deutschen Sprahe.* Vuisshaya Skola, Mosku.

Church, K. eta Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Church, K., Hanks, P., eta Hindle, D. nad Gale, W. (1991). Using statistics in lexical analysis. Zernik, (ed.), *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, 115–164. Lawrence Erlbaum Associates.

Cimiano, P. (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications* 27. lib. Springer-Verlag New York, Inc.

Contreras, J. M. eta Suñer, A. l. (2004). Los procesos de lexicalización. Zabala, I., Pérez Gaztelu, E., eta García, L.(ed.), *Las fronteras de la composición en lenguas románicas y en vasco*, 47–108. Universidad de Deusto, Servicio de Publicaciones, Donostia.

Corpas Pastor, G. (1996). *Manual de Fraseología Española.* Gredos, Madrid.

Corpas Pastor, G. (2001). En torno al concepto de colocación. *EUSKERA*, 46:89–108.

Coseriu, E. (1977). *Principios de Semántica Estructural.* Editorial Gredos.

Cowie, A. (1998a). *Phraseology: Theory, Analysis, and Applications.* Oxford University Press, USA.

Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics*, 2(3):223–235.

Cowie, A. P. (1986). Collocational dictionaries - A comparative view. *Proceedings of the Fourth Joint Anglo-Soviet Seminar on English Studies. The British Council, London*, 61–69.

Cowie, A. P. (1988). Stable and creative aspects of vocabulary use. *Vocabulary and Language Teaching*, 126–39.

Cowie, A. P. (1998b). Phraseological dictionaries: some east-west comparisons. Cowie, A., (ed.), *Phraseology: Theory, Analysis, and Applications*, 209–228. Oxford University Press, USA.

Cowie, A. P. eta Howarth, P. (1996). Phraseological competence and written proficiency. *British Studies In Applied Linguistics*, 11:80–93.

Cowie, A. P., Mackin, R., eta McCaig, I. (1983). *Oxford Dictionary of Current Idiomatic English*. Oxford University Press.

Cruse, D. (1986). *Lexical Semantics*. Cambridge University Press.

Curran, J. R. eta Moens, M. (2002). Improvements in automatic thesaurus extraction. *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition* 9. lib., 59–66. Association for Computational Linguistics.

Dagan, I. (2000). Contextual word similarity. Robert, D., Moisl, H. L., eta Somers, H. L.(ed.), *Handbook of Natural Language Processing. Marcel Dekker*, 459–475. Marcel Dekker Inc, New York, NY.

Daille, B. (1994). *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. Doktorego-tesia, Université Paris 7, Frantzia.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., eta Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Dias, G. (2003). Multiword unit hybrid extraction. *Proceedings of the ACL 2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 41–48. Association for Computational Linguistics.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Elhuyar (2006). *Elhuyar Hiztegia. Euskara/Gaztelania - Castellano/Vasco*. Elhuyar Fundazioa, Usurbil.

Erman, B. eta Warren, B. (2000). The idiom principle and the open choice principle. *Text*, 20(1):29–62.

Esnal, P. (2001). "Ortik eta emendik": euskal lokuzioak eta fraseologia baino ere haratago. *Euskera: Euskaltzaindiaren lan eta agiriak*, 46(1):137–144.

Etxebarria, J. R. eta Bilbao, X. (2012). Magnitude fisikoekin eratzen diren <izena + aditza> motako kolokazioak euskaraz, bost magnituderen kasuan. *Uztaro: giza eta gizarte-zientzien aldizkaria*, 82:55–81.

Etxepare, R. (2003). Valency and argument structure in the Basque verb. Hualde, J. eta Ortiz de Urbina, J.(ed.), *A grammar of Basque*, 363–425. Mouton de Gruyter, Berlin.

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Doktorego-tesia, University of Stuttgart.

Evert, S. (2008). Corpora and collocations. Lüdeling, A. eta Kytö, M.(ed.), *Corpus Linguistics. An International Handbook* 29. lib., 1212–1247. De Gruyter Mouton.

Evert, S. eta Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 188–195. Association for Computational Linguistics.

Evert, S. eta Krenn, B. (2005a). Exploratory collocation extraction. *PHRASEOLOGY 2005 - The many faces of Phraseology*. Université catholiquede Louvain, Belgika.

Evert, S. eta Krenn, B. (2005b). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, 19(4):450–466.

Ezeiza, N. (2002). *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile sintaktiko sendo eta malgua*. Doktorego-tesia, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU, Donostia.

Ezeiza, N., Alegria, I., Arriola, J. M., Urizar, R., eta Aduriz, I. (1998). Combining stochastic and rule-based methods for disambiguation in agglutinative languages. *Proceedings of the 17th International Conference on Computational Linguistics* 1. lib., 380–384. Association for Computational Linguistics.

Fazly, A., Cook, P., eta Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Fazly, A. eta Stevenson, S. (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 9–16. Association for Computational Linguistics.

Fernandez, B. (1997). *Egiturazko kasuen erkaketa euskaraz*. Doktorego-tesia, Euskal Herriko Unibertsitatea.

Fernandez, I. (2012). *Euskarazko Entitate-Izenak: identifikazioa, sailkapena, itzulpena eta desanbiguazioa*. Doktorego-tesia, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU, Donostia.

Fernando, C. eta Flavell, R. (1981). *On Idiom: Critical Views and Perspectives*. University of Exeter.

Fillmore, C. J. (1979). On fluency. Kempler, D. eta Wang, W. S. Y.(ed.), *Individual Differences in Language Ability and Language Behavior*, 85–101. Academic Press, New York.

Fillmore, C. J., Kay, P., eta O'connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language*, 501–538.

Firth, J. (1957). *Papers in Linguistics 1934–1951*. Oxford University Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Fontenelle, T. (1998). Discovering significant lexical functions in dictionary entries. Cowie, A., (ed.), *Phraseology: Theory, Analysis, and Applications*, 189–207. Oxford University Press, USA.

Fraser, B. (1970). Idioms within a transformational grammar. *Foundations of Language*, 6(1):22–42.

Fredricks, G. A. eta Nelsen, R. B. (2007). On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. *Journal of Statistical Planning and Inference*, 137(7):2143–2150.

Frontini, F., Quochi, V., eta Rubino, F. (2012). Automatic creation of quality multi-word lexica from noisy text data. *Proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data*, Mumbai, India.

Fung, P. eta Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. *Proceedings of the 17th International Conference on Computational Linguistics* 1. lib., 414–420. Association for Computational Linguistics.

Garate, G. (1998). *Erdarakadak: Euskaraz ongi mintzatzeko hiztegia. Zazpi probintzietako adibideak*. Gero Euskal Liburuak - Ediciones Mensajero, Bilbo.

Garate, G. (2003). *Atsotitzak*. Bilbao Bizkaia Kutxa Fundazioa, Bilbo.

Garrão, M., Oliveira, C., de Freitas, M. C., eta Dias, M. C. (2006). Corpus-based compositionality. *Proceedings of the 7th International Conference on Computational Processing of the Portuguese Language*, 268–271, Berlin, Heidelberg. Springer-Verlag.

Garrido, G. eta Peñas, A. (2011). Detecting compositionality using semantic vector space models based on syntactic context: shared task system description. *Proceedings of the Workshop on Distributional Semantics and Compositionality*, 43–47. Association for Computational Linguistics.

Gayen, V. eta Sarkar, K. (2013). Automatic identification of Bengali noun-noun compounds using Random Forest. *Proceedings of the 9th Workshop on Multiword Expressions* 13. lib., 64–72, Atlanta, Georgia, AEB. Association for Computational Linguistics.

Gibbons, J. D. (1993). *Nonparametric Measures of Association* 91. lib. Sage University Paper Series on Quantitative Applications in the Social Sciences, Newbury Park, CA, AEB.

Gibbs, R. W., Bogdanovich, J. M., Sykes, J. R., eta Barr, D. J. (1997). Metaphor in idiom comprehension. *Journal of Memory and Language*, 37(2):141–154.

Gilpin, A. R. (1993). Table for conversion of Kendall's tau to Spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educational and Psychological Measurement*, 53(1):87–92.

Gläser, R. (1988). The grading of idiomaticity as a presupposition for a taxonomy of idioms. *Understanding the lexicon: Meaning, sense and world knowledge in lexical semantics*, 264–279.

Gläser, R. (1998). The stylistic potential of phraseological units in the light of genre analysis. Cowie, A., (ed.), *Phraseology: Theory, Analysis, and Applications*, 125–143. Oxford University Press, USA.

Goldman, J.-P., Nerima, L., eta Wehrli, E. (2001). Collocation extraction using a syntactic parser. *Proceedings of the ACL Workshop on Collocations*, 61–66.

González Rey, M. I. (1998). Estudio de idiomaticidad en las unidades fraseológicas. Wotjak, G., (ed.), *Estudios de fraseología y fraseografía del español actual*, 57–74. Lingüística Iberoamericana.

Granger, S. eta Paquot, M. (2008). Disentangling the phraseological web. Granger, S. eta Meunier, F.(ed.), *Phraseology: An Interdisciplinary Perspective*, 27–50. John Benjamins Publishing Co.

Grefenstette, G., Heid, U., eta Fontenelle, T. (1996). The DECIDE project: Multilingual collocation extraction. *Proceedings of the 6th International Congress of Lexicography - EURALEX '96*, 93–108, Göteborg.

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. Granger, S. eta Meunier, F.(ed.), *Phraseology: An Interdisciplinary Perspective*, 3–25. John Benjamins Publishing, Amsterdam/Filadelfia.

Grimshaw, J. eta Mester, A. (1988). Light verbs and θ-marking. *Linguistic Inquiry*, 205–232.

Guevara, E. (2010). A regression model of adjective-noun compositionality in distributional semantics. *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*, 33–37. Association for Computational Linguistics.

Gurrutxaga, A. (2002). Euskal hiztegietako eta corpusetako izen+izenondo kolokazio batzuen azterketa konparatiboa: kolokazioak corpusetatik erauzteak ekar dezakeen hobekuntza. HIZTEK graduatu ondoko ikastaroa. Proiektua. Euskal Herriko Unibertsitatea.

Gurrutxaga, A. (2003). Hitz anitzeko unitateen tratamendua hiztegigintzan: Elhuyarren hiztegigintzako datu-baserako proposamenak. HIZTEK graduatu ondoko masterra. Proiektua. Euskal Herriko Unibertsitatea.

Gurrutxaga, A. eta Alegria, I. (2012). Measuring the compositionality of NV expressions in basque by means of distributional similarity techniques.

Proceedings of the 8th international Conference on Language Resources and Evaluation - LREC 2012, 2389–2394, Istanbul.

Gwet, K. L. (2012). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters.* Advanced Analytics Press.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., eta Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Hanks, P. (2004). The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, 17(3):245–274.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Harris, Z. S. (1970). *Papers in Structural and Transformational Linguistics*. D. Reidel, Dordrecht, Herbehereak.

Hauser, M. D., Chomsky, N., eta Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579.

Hausmann, F. J. (1989). Le dictionnaire de collocations. *Wörterbücher, Dictionaries*, 1:1010–1019.

Heid, U. (1994). On ways words work together–Research topics in lexical combinatorics. *Proceedings of the 6th International Congress of Lexicography - EURALEX '94*, 226–257.

Heid, U. (1998). Towards a corpus-based dictionary of German noun-verb collocations. *Proceedings of the 8th International Congress of Lexicography - EURALEX '98*, 301–312.

Heid, U. (2008). Computational phraseology. An overview. Granger, S. eta Meunier, F.(ed.), *Phraseology: An Interdisciplinary Perspective*, 337–360. John Benjamins Publishing, Amsterdam/Filadelfia.

Hoang, H. H., Kim, S.Ñ., eta Kan, M.-Y. (2009). A re-examination of lexical association measures. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, 31–39. Association for Computational Linguistics.

Howarth, P. (1998). The phraseology of learners' academic writing. Cowie, A., (ed.), *Phraseology: Theory, Analysis, and Applications*, 161–186. Oxford University Press, USA.

Howarth, P. A. (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making* 75. lib. Walter de Gruyter.

Inkpen, D. Z. eta Hirst, G. (2002). Acquiring collocations for lexical choice between near-synonyms. *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition* 9. lib., 67–76. Association for Computational Linguistics.

Irsula, J. (1994). Entre el verbo y el sustantivo, ¿quién rige a quién? El verbo en las colocaciones sustantivo-verbales. *Verbo e estruturas frásicas/Colóquio Internacional de Linguística Hispânica*, 277–286.

Izagirre, K. (1981). *Euskal Lokuzioak. Espainolezko eta frantsesezko gidazerrendarekin.* Hordago, Donostia.

Jackendoff, R. (1995). The boundaries of the lexicon. *Idioms: Structural and Psychological Perspectives*, 133–165.

Kaltzakorta, X. (2001). Euskal fraseologia: historia, oinarriak. *Euskera: Euskaltzaindiaren lan eta agiriak*, 46(1):73–87.

Katz, G. eta Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using Latent Semantic Analysis. *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 12–19. Association for Computational Linguistics.

Keller, F. (2003). Connectionist and statistical language processing. Computerlinguistik, Universität des Saarlandes. http://www. coli. unisaarland. de/˜ crocker/courses/learning/lecture10.pdf.

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics.* Longman.

Keshavarz, M. H. eta Salimi, H. (2007). Collocational competence and cloze test performance: A study of iranian efl learners. *International Journal of Applied Linguistics*, 17(1):81–92.

Kilgarriff, A. (1996). Which words are particularly characteristic of a text? A survey of statistical approaches. *Language Engineering for Document Analysis and Recognition*, 33–40.

Kim, S.Ñ. eta Baldwin, T. (2007). Detecting compositionality of English verb-particle constructions using semantic similarity. *Proceedings of the 7th Meeting of the Pacific Association for Computational Linguistics (PACLING 2007)*, 40–48.

Klinkenberg, J.-M. (1990). *Le sens rhétorique: essais de sémantique littéraire*. Éd. du GREF, Toronto.

Kobyliński, Ł. eta Przepiórkowski, A. (2008). Definition extraction with balanced random forests. *Advances in Natural Language Processing*, 237–247. Springer.

Koike, K. (1998). Algunas observaciones sobre colocaciones sustantivo-verbales. Wotjak, G., (ed.), *Estudios de fraseología y fraseografía del español actual*, 245–256. Lingüística Iberoamericana.

Kontonatsios, G., Korkontzelos, I., Tsujii, J., eta Ananiadou, S. (2013). Using a Random Forest Classifier to recognise translations of biomedical terms across languages. *ACL 2013*, 95.

Korhonen, A., Krymolowski, Y., eta Briscoe, T. (2006). A large subcategorization lexicon for natural language processing applications. *Proceedings of the 5th International Conference on Language Resources and Evaluation - LREC 2006*, 1015–1020, Genoa.

Korkontzelos, I. (2010). *Unsupervised Learning of Multiword Expressions*. Doktorego-tesia, University of York.

Korkontzelos, I. eta Manandhar, S. (2009). Detecting compositionality in multi-word expressions. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 65–68. Association for Computational Linguistics.

Kotsiantis, S. B. (2007). Supervised Machine Learning: a review of classification techniques. *Informatica*, 31(3):249–?268.

Krcmár, L., Jezek, K., eta Pecina, P. (2013). Determining compositionality of word expressions using word space models. *Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013) NAACL HLT 2013*, 13:42–50.

Krenn, B. (2000). *The usual suspects: Data-oriented models for identification and representation of lexical collocations.* Doktorego-tesia, Universität des Saarlandes, Saarbrücken.

Krenn, B. (2004). Manual zur Identifikation von Funktionsverbgefügen und figurativen Ausdrücken in PP-Verb-Listen. *Austrian Research Institute for Artificial Intelligence.*

Krenn, B. (2008). Description of evaluation resource – German PP-verb data. *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, 7–10.

Krenn, B. eta Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. *Proceedings of the ACL Workshop on Collocations*, 39–46.

Krenn, B., Evert, S., eta Zinsmeister, H. (2004). Determining intercoder agreement for a collocation identification task. *Proceedings of KONVENS*, 89–96.

Kullback, S. eta Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 79–86.

Laka, I. (1993). Unergatives that assign ergative, unaccusatives that assign accusative. *MIT Working Papers in Linguistics*, 18:149–172.

Lakarra, J. (1996). *Refranes y sentencias (1596): ikerketak eta edizioa.* Euskaltzaindia.

Landis, J. R. eta Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Lee, L. (1999). Measures of distributional similarity. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, 25–32. Association for Computational Linguistics.

Levin, B. C. (1983). *On the Nature of Ergativity.* Doktorego-tesia, Massachusetts Institute of Technology.

Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 64–71. Association for Computational Linguistics.

Lin, D. (1998). Automatic retrieval and clustering of similar words. *Proceedings of the 17th International Conference on Computational Linguistics* 2. lib., 768–774. Association for Computational Linguistics.

Lin, D. (1999). Automatic identification of non-compositional phrases. *Proceedings of the 37th Annual Meeting of the ACL*, 317–324. Association for Computational Linguistics.

Lin, D. eta Wu, X. (2009). Phrase clustering for discriminative learning. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* 2. lib., 1030–1038. Association for Computational Linguistics.

Lin, J., Li, S., eta Cai, Y. (2008). A new collocation extraction method combining multiple association measures. *Proceedings of International Conference on Machine Learning and Cybernetics* 1. lib., 12–17. IEEE.

Linares, M. A. M. (2006). Palabra y lexía. *Biblioteca de recursos electrónicos de humanidades. E-xcelence.*

Lipka, L., Handl, S., eta Falkner, W. (2004). Lexicalization and institutionalization. The state of the art in 2004. *SKASE Journal of Theoretical Lingsuitics*, 1:2–19.

Lopez de Lacalle, O. (2009). *Domain-Specific Word Sense Disambiguation.* Doktorego-tesia, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU, Donostia.

Lorente, M. (2001). Altres elements léxics. Solà, J., (ed.), *Gramàtica del Català Contemporari*, 831–888. Empúries, Bartzelona.

Lund, K. eta Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

Makkai, A. (1972). *Idiom Structure in English* 48. lib. Mouton The Hague.

Maldonado-Guerra, A. eta Emms, M. (2011). Measuring the compositionality of collocations via word co-occurrence vectors: Shared task system description. *Proceedings of the Workshop on Distributional Semantics and Compositionality*, 48–53. Association for Computational Linguistics.

Manning, C. D. eta Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Marrero, M., Sanchez-Cuadrado, S., Lara, J. M., eta Andreadakis, G. (2009). Evaluation of named entity extraction systems. *Advances in Computational Linguistics, Research in Computing Science*, 41:47–58.

Martínez, A. (1996). Syntactic evidence in favour of degrees of incorporation in [*n*+*egin*] constructions. Eskuizkribua.

Matsumoto, Y. eta Utsuro, T. (2000). Lexical knowledge acquisition. Dale, H. M. eta Somers, H.(ed.), *Handbook of Natural Language Processing*, 563–610. Marcel Dekker.

McCarthy, D. (2008). Lexical substitution as a framework for multiword evaluation. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC '08)*.

McCarthy, D., Keller, B., eta Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. *Proceedings of the ACL 2003 Workshop on Multiword Expressions: analysis, acquisition and treatment* 18. lib., 73–80. Association for Computational Linguistics.

McKeown, K. R. eta Radev, D. R. (2000). Collocations. Robert, D., Moisl, H. L., eta Somers, H. L.(ed.), *Handbook of Natural Language Processing. Marcel Dekker*, 507–521. Marcel Dekker Inc, New York, NY.

Mel'čuk, I. (1988). Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, 1(3):165–188.

Mel'čuk, I., Arbatchewsky-Jumarie, N., Elnitsky, L., Iordanskaja, L., Lessard, A., Dagenais, L., Lefebvre, M.-N., eta Mantha, S. (1984). Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques i. *Les Presses de l'Université de Montréal*.

Mel'cuk, I. eta Wanner, L. (1994). Towards an efficient representation of restricted lexical cooccurrence. *Proceedings of the 6th International Congress of Lexicography - EURALEX '94* 94. lib.

Mel'čuk, I. A. eta Polguere, A. (1987). A formal lexicon in the Meaning-Text Theory (or how to do lexica with words). *Computational linguistics*, 13(3-4):261–275.

Mel'čuk, I. A. eta Polguère, A. (2007). *Lexique actif du français: l'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. De Boeck.

Mel'čuk, I. (1998). Collocations and lexical functions. Cowie, A., (ed.), *Phraseology: Theory, Analysis, and Applications*, 23–53. Oxford University Press, USA.

Mitchell, J. eta Lapata, M. (2008). Vector-based models of semantic composition. *Proceedings of ACL-08: HLT*, 236–244, Columbus, Ohio, AEB. Association for Computational Linguistics.

Mokoroa, J. (1990). *Ortik eta emendik. Repertorio de locuciones del habla popular vasca, oral y escrita, en sus diversas variedades*. Labayru. Eusko Jaurlaritza, Bilbo.

Montero Martínez, S. (2002). *Estructuración conceptual y formalización terminográfica de frasemas en el subdominio de la oncología*. Doktoregotesia, Universidad de Valladolid.

Moon, R. (1998a). *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Clarendon Press Oxford.

Moon, R. (1998b). Frequencies and forms of phrasal lexemes in English. Cowie, A., (ed.), *Phraseology: Theory, Analysis, and Applications*, 79–100. Oxford University Press, USA.

Nunberg, G., Sag, I. A., eta Wasow, T. (1994). Idioms. *Language*, 491–538.

Odriozola, J. (2010). Euskararen aditz-unitate fraseologikoen deskribapena. Unibertsitateko katedra plazarako lehiaketa. Zientzia eta Teknologia fakultatea.

Omazić, M. (2008). Processing of idioms and idiom modifications. A view from cognitive linguistics. Granger, S. eta Meunier, F.(ed.), *Phraseology: An Interdisciplinary Perspective*, 67–79. John Benjamins Publishing, Amsterdam/Filadelfia.

Ooi, V. B. eta Coi, V. (1998). *Computer Corpus Lexicography*. Edinburgh University Press.

Ortiz de Urbina, J. (1989). *Parameters in the grammar of Basque: A GB approach to Basque syntax*. Foris Publications USA, Dordrecht.

Otegi, A. (2012). *Hedapena informazioaren berreskurapenean: hitzen adiera-desanbiguazioaren eta antzekotasun semantikoaren ekarpenak*. Doktoregotesia, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU, Donostia.

Oyharçabal, B. (1994). Contribution de la comparaison typologique à une analyse des rapports ergativité-(in) transitivité en basque'. *La langue basque parmi les autres (influences et comparaison), Izpegi, Saint-Etienne-de-Baigorry*, 115–148.

Oyharçabal, B. (2006). Basque light verb constructions. *Anuario del Seminario de Filología Vasca Julio de Urquijo. R. L. Trasken oroitzapenetan ikerketak euskalaritzaz eta hizkuntzalaritza historikoaz*, 40(1–2):787–806.

Padó, S. eta Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Partee, B. (1995). Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360.

Pawley, A. eta Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and Communication*, 191:225.

Pearce, D. (2001). Synonymy in collocation extraction. *Proceedings of the NAACL'01 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations.*, 41–46, Pittsburgh, PA, AEB. Association for Computational Linguistics.

Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. *In Proceedings of the 3th International Conference on Language Resources and Evaluation - LREC 2002*, 1530–1536, Las Palmas de Gran Canaria.

Pecina, P. (2005). An extensive empirical study of collocation extraction methods. *Proceedings of the ACL Student Research Workshop*, 13–18. Association for Computational Linguistics.

Pecina, P. (2009). *Lexical Association Measures: Collocation Extraction* 4. lib. , *Studies in Computational and Theoretical Linguistics*. Institute of Formal and Applied Linguistics, Charles University, Praga, Txekia.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1):137–158.

Pecina, P. eta Schlesinger, P. (2006). Combining association measures for collocation extraction. *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, 651–658. Association for Computational Linguistics.

Pedersen, T. (1996). Fishing for exactness. *Proceedings of the South-Central SAS Users Group Conference*.

Pedersen, T., Banerjee, S., McInnes, B. T., Kohli, S., Joshi, M., eta Liu, Y. (2011). The Ngram Statistics Package (text::NSP): A flexible tool for identifying Ngrams, collocations, and word associations. *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, 131–133. Association for Computational Linguistics.

Pérez, J. (2006). *Árboles consolidados: construcción de un árbol de clasificación basado en múltiples submuestras sin renunciar a la explicación*. Doktorego-tesia, Informatika Fakultatea, Euskal Herriko Unibertsitatea, Donostia.

Pinker, S. (1994). *The Language Instinct*. New York: William Morrow and Company.

Pociello, E., Agirre, E., eta Aldezabal, I. (2011). Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45(2):121–142.

Polguère, A. (2000). Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. *Proceedings of EURALEX 2000*, 517–527, Stuttgart.

Rafel, J. (2004). Los pedicados complejos en español. Zabala, I., Pérez Gaztelu, E., eta García, L.(ed.), *Las fronteras de la composición en lenguas románicas y en vasco*, 393–443. Universidad de Deusto, Servicio de Publicaciones, Donostia.

Ramisch, C. (2012). *A generic and open framework for multiword expressions treatment: from acquisition to applications*. Doktorego-tesia, University of Grenoble (France) and Federal University of Rio Grande do Sul (Brazil), Grenoble, France.

Reddy, S., McCarthy, D., Manandhar, S., eta Gella, S. (2011). Exemplar-based word-space model for compositionality detection: Shared task system description. *Proceedings of the Workshop on Distributional Semantics and Compositionality*, 54–60. Association for Computational Linguistics.

Rennie, J. D., Shih, L., Teevan, J., Karger, D. R., et al. (2003). Tackling the poor assumptions of naive bayes text classifiers. *Proceedings of ICML 2003* 3. lib., 616–623. Washington DC.

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1):127–159.

Roberts, M. H. (1944). The science of idiom: a method of inquiry into the cognitive design of language. *Publications of the Modern Language Association of America*, 291–306.

Rodriguez, S. eta Murga, F. G. (2003). *IZEN+ EGIN* predikatuak euskaraz. *Euskal Gramatikari eta literaturari buruzko Jardunaldiak XXI. mendearen atarian (I-II)*, 417–436. Euskaltzaindia.

Ruiz Gurillo, L. (1998). Una clasificación no discreta de las unidades fraseológicas del español. Wotjak, G., (ed.), *Estudios de fraseología y fraseografía del español actual*, 13–37. Lingüística Iberoamericana.

Sag, I., Baldwin, T., Bond, F., Copestake, A., eta Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing - CICLING 2002*, 1–15, Londres. Springer-Verlag.

Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Doktorego-tesia, Stockholm.

Salton, G., Wong, A., eta Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Salvador, V. (2000). Idiomaticitat i discurs prefabricat. Salvador, V. eta Piquer, A.(ed.), *El discurs prefabricat*, 19–31. Universitat Jaume I.

Saralegi, X., San Vicente, I., eta Gurrutxaga, A. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain. *Proceedings of the 6th International Conference on Language Resources and Evaluation -LREC 2008 - Building and using Comparable Corpora workshop*, 27–32.

Sarasola, I. (1996). *Euskal Hiztegia*. Kutxa Fundazioa / Fundación Kutxa, Donostia.

Sarasola, I. (1997). *Euskara batuaren ajeak*. Alberdania.

Schmitt, N. eta Carter, R. (2004). Formulaic sequences in action. Schmitt, N., (ed.), *Formulaic Sequences: Acquisition, Processing and Use*, 1–22. John Benjamins Publishing.

Schone, P. eta Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing - EMNLP 2001*, 100–108, Pittsburgh, PA, AEB.

Schütze, H. (1993). Word space. Hanson, S. J., Cowan, J. D., eta Giles, C. L.(ed.), *Advances in Neural Information Processing Systems 5*, 895–902. Morgan Kaufmann Publishers Inc.

Schütze, H. eta Pedersen, J. O. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, 33(3):307–318.

Seretan, V. (2011). *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Springer, Dordrecht.

Seretan, V. (2013). On collocations and their interaction with parsing and translation. *Informatics*, 1(1):11–31.

Seretan, V. eta Wehrli, E. (2006). Accurate collocation extraction using a multilingual parser. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 953–960. Association for Computational Linguistics.

Sinclair, J. (1991). *Corpus, Concordance, Collocation* 1. lib. Oxford University Press Oxford.

Sinclair, J. (1996). The search for units of meaning. *Textus*, 9(1):75–106.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Smadja, F., McKeown, K. R., eta Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1):1–38.

Specia, L. (2005). Knowledge sources for disambiguating highly ambiguous verbs in machine translation. *Proceedings of the 17th European Summer School in Logic, Language and Information, ESSLLI-2005*, Edinburgh.

Street, L., Michalov, N., Silverstein, R., Reynolds, M., Ruela, L., Flowers, F., Talucci, A., Pereira, P., Morgon, G., Siegel, S., Barousse, M., Anderson, A., Carroll, T., eta Feldman, A. (2010). Like finding a needle in a haystack: Annotating the American National Corpus for idiomatic expressions. *Proceedings of the 7th International Conference on Language Resources and Evaluations - LREC 2010*, Valletta, Malta.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell Publishers.

Svensson, M. H. (2008). A very complex criterion of fixedness: Noncompositionality. Granger, S. eta Meunier, F.(ed.), *Phraseology: An Interdisciplinary Perspective*, 81–93. John Benjamins Publishing, Amsterdam/Filadelfia.

Thanopoulos, A., Fakotakis, N., eta Kokkinakis, G. (2002). Comparative evaluation of collocation extraction metrics. *Proceedings of the 3rd International Conference on Language Resources and Evaluation - LREC 2002* 2. lib., 620–625.

Tristá Pérez, A. M. (1998). La fraseología y la fraseografía. Wotjak, G., (ed.), *Estudios de fraseología y fraseografía del español actual*, 297–306. Lingüística Iberoamericana.

Turney, P. D., Pantel, P., et al. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Turpin, A. eta Scholer, F. (2006). User performance versus precision measures for simple search tasks. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 11–18. ACM.

Tutin, A. (2005). Le dictionnaire de collocations est-il indispensable? *Revue française de linguistique appliquée*, 10(2):31–48.

Tutin, A. eta Grossmann, F. (2002). Collocations régulières et irrégulières: esquisse de typologie du phénomène collocatif. *Revue française de linguistique appliquée*, VII:7–25.

Uribe-Etxebarria, M. (1989). *On noun incorporation in Basque and some of its consequences in the phrase structure*. Ms, University of Connecticut, Storrs, CT, AEB.

Urizar, R. (2012). *Euskal lokuzioen tratamendu konputazionala.* Doktorego-tesia, Informatika Fakultatea, UPV/EHU, Donostia.

Van de Cruys, T. eta Moirón, B. (2007). Semantics-based multiword expression extraction. *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 25–32. Association for Computational Linguistics.

Van Rijsbergen, C. J. (1979). *Information Retrieval.* Butterworths, Londres.

Venkatapathy, S. eta Joshi, A. K. (2005). Measuring the relative compositionality of verb-noun (VN) collocations by integrating features. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 899–906. Association for Computational Linguistics.

Vickrey, D., Biewald, L., Teyssier, M., eta Koller, D. (2005). Word-sense disambiguation for machine translation. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 771–778. Association for Computational Linguistics.

Viegas, E. eta Bouillon, P. (1994). Semantic lexicons: the cornerstone for lexical choice in natural language generation. *Proceedings of the Seventh International Workshop on Natural Language Generation*, 91–98. Association for Computational Linguistics.

Villar, Z. S. (2011). Alemanetik euskarara itzulitako unitate fraseologikoen azterketarako jarraibideak. *Senez: itzulpen aldizkaria*, 41:125–139.

Vincze, V. (2012). Light verb constructions in the SzegedParalellFX English-Hungarian Parallel Corpus. *Proceedings of the 8th International Conference on Language Resources and Evaluations - LREC 2012*, 2381–2388.

Vincze, V. (2013). *Semi-compositional noun+verb constructions: Theoretical questions and computational linguistic analyses.* Doktorego-tesia, LAP LAMBERT Academic Publishing.

Vinogradov, V. (1947). Ob osnovnuikh tipakh frazeologicheskikh edinits v russkom yazuike. *A .A. Shakhmatov, 1864-1920. Sbornik statey i materialov*, 339–64. Mosku: Nauka.

Völker, J., Vrandečić, D., Sure, Y., eta Hotho, A. (2007). Learning disjointness. *The Semantic Web: Research and Applications*, 175–189. Springer.

Von Polenz, P. (1963). *Funktionsverben im heutigen Deutsch: Sprache in der rationalisierten Welt.* Pädogogischer Verlag Schwann.

Warren, B. (2005). A model of idiomaticity. *Nordic Journal of English Studies*, 4(1):35–54.

Weeds, J. E. (2003). *Measures and Applications of Lexical Distributional Similarity.* Doktorego-tesia, University of Sussex.

Wehrli, E. (2007). Fips, a deep linguistic multilingual parser. *Proceedings of the Workshop on Deep Linguistic Processing*, 120–127. Association for Computational Linguistics.

Wermter, J. eta Hahn, U. (2004). Collocation extraction based on modifiability statistics. *Proceedings of the 20th International Conference on Computational Linguistics*, 980. Association for Computational Linguistics.

Wermter, J. eta Hahn, U. (2006). You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 785–792. Association for Computational Linguistics.

Wiktorsson, M. (2003). *Learning Idiomaticity : A Corpus-Based Study of Idiomatic Expressions in Learners' Written Production.* Doktorego-tesia, Lund University, Lund, Suedia.

Witten, I. H. eta Frank, E. (2005). *Data Mining. Practical Machine Learning Tools and Techniques with Java Inplementations.* Morgan Kaufmann Publishers.

Wotjak, G. (1998). Reflexiones acerca de construcciones verbo-nominales funcionales. Wotjak, G., (ed.), *Estudios de fraseología y fraseografía del español actual*, 257–280. Lingüística Iberoamericana.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, 21(4):463–489.

Wulff, S. (2008). *Rethinking Idiomaticity.* Corpus and Discourse. Continuum International Publishing Group Ltd, New York.

Yates, F. (1984). Tests of significance for 2× 2 contingency tables. *Journal of the Royal Statistical Society. Series A (General)*, 426–463.

Zabala, I. (2004). Los predicados complejos en vasco. Zabala, I., Pérez Gaztelu, E., eta García, L.(ed.), *Las fronteras de la composición en lenguas románicas y en vasco*, 445–567. Universidad de Deusto, Servicio de Publicaciones, Donostia.

Zabell, S. L. (1989). The rule of succession. *Erkenntnis*, 31(2-3):283–321.

Zamarripa, P. (1913). *Manual del vascófilo.* Imprenta y Ene. de José A. de Lerchundi, Bilbo.

Zaninello, A. eta Nissim, M. (2010). Creation of lexical resources for a characterisation of multiword expressions in Italian. *Proceedings of the 7th International Conference on Language Resources and Evaluations - LREC 2010.*

Zavala, A. (1985). *Esaera zaarren bilduma berria (1/2).* Auspoa, Tolosa.

Zelaia, A., Baragaña, I., eta Yurramendi, Y. (2011). LSAren oinarri matematikoa. *EKAIA Euskal Herriko Unibertsitateko Zientzi eta Teknologi Aldizkaria*, 17:85–105.

Zgusta, L. (1971). *Manual of Lexicography* 39. lib. Walter de Gruyter.

Zhu, M. (2004). *Recall, precision and average precision.* Working paper 2004-09, Department of Statistics and Actuarial Science, University of Waterloo.

Zuluaga, A. (1980). *Introducción al estudio de las expresiones fijas.* Frankfurt: P.D. Lang.

Lengoaia eta Sistema Informatikoak Saila

Informatika Fakultatea

Idiomatikotasunaren karakterizazio automatikoa: `izena+aditza` konbinazioak

Eranskinak

Antton Gurrutxaga Hernaizek

Informatikan Doktore titulua eskuratzeko aurkezturiko

TESI-TXOSTENA

Donostia, 2014ko ekaina

A. ERANSKINA

Ebaluazio-erreferentzia

Taulan erabilitako laburtzapenenak

Kasu-marka

ABL	ablatiboa
ABS	absolutiboa
ABU	muga-adlatiboa
ABZ	hurbiltze-adlatiboa
ALA	adlatiboa
DAT	datiboa
DES	destinatiboa
ERG	ergatiboa
GEL	genitibo leku-denborazkoa
GEN	genitibo edutezkoa
INE	inesiboa
INS	instrumentala
MOT	motibatiboa
PAR	partitiboa
PRO	prolatiboa
SOZ	soziatiboa

Mugatasuna eta numeroa

MG	Mugagabea
NUMS	Mugatu singularra
NUMP	Mugatu plurala
PH	Plural hurbila

UF-kategoriak

id	esapide idiomatikoa
col	kolokazioa
free	konbinazio librea

Bigrama normalizatua					**Sailkapena**	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
abenturak	abentura	ABS	NUMP	**kontatu**	free	free
abiadura	abiadura	ABS	NUMS	**moteldu**	col	col
abiapuntura	abiapuntu	ALA	NUMS	**itzuli**	free	free
abiapuntua	abiapuntu	ABS	NUMS	**izan**	col	col
abilezia	abilezia	ABS	NUMS	**erakutsi**	free	free
adarretatik	adar	ABL	NUMP	**heldu**	id	id
adarra	adar	ABS	NUMS	**jo**	id	id
adibide	adibide	ABS	MG	**izan**	free	free
aditzerat	aditze	ALA	NUMS	**eman**	col	col
administrazioa	administrazio	ABS	NUMS	**euskaldundu**	free	free
adostasunera	adostasun	ALA	NUMS	**heldu**	col	col
adostasuna	adostasun	ABS	NUMS	**lortu**	col	col
agerraldia	agerraldi	ABS	NUMS	**egin**	col	col
aginduei	agindu	DAT	NUMP	**jarraitu**	free	free
agindua	agindu	ABS	NUMS	**jaso**	col	col
agintaldia	agintaldi	ABS	NUMS	**amaitu**	free	free
agintean	aginte	INE	NUMS	**egon**	col	col
agintean	aginte	INE	NUMS	**izan**	free	free
agiriak	agiri	ABS	NUMP	**faltsutu**	free	free
agiria	agiri	ABS	NUMS	**kaleratu**	col	col
agiria	agiri	ABS	NUMS	**sinatu**	free	free
ahaleginetan	ahalegin	INE	NUMP	**ibili**	col	col
ahal	ahal	ABS	MG	**izan**	col	id
ahal	ahal	ABS	MG	**ukan**	col	id
ahotsa	ahots	ABS	NUMS	**isilarazi**	free	free
ahotsa	ahots	ABS	NUMS	**izan**	id	id
aipamena	aipamen	ABS	NUMS	**egin**	col	col
aireportua	aireportu	ABS	NUMS	**handitu**	free	free
aireportura	aireportu	ALA	NUMS	**iritsi**	free	free
aireportuan	aireportu	INE	NUMS	**lurreratu**	free	free
aitzindaria	aitzindari	ABS	NUMS	**izan**	free	free
aitzinetik	aitzin	ABL	NUMS	**pasatu**	free	free
akatsak	akats	ABS	NUMP	**egin**	col	col
akordioa	akordio	ABS	NUMS	**izan**	free	free
akzioak	akzio	ABS	NUMP	**salerosi**	free	free
albiste	albiste	ABS	MG	**izan**	col	col
alboan	albo	INE	NUMS	**egon**	free	free
albotik	albo	ABL	NUMS	**igaro**	free	free
alboan	albo	INE	NUMS	**izan**	id	id
albora	albo	ALA	NUMS	**utzi**	id	id
aldaketak	aldaketa	ABS	NUMP	**egin**	col	col
aldaketak	aldaketa	ABS	NUMP	**ekarri**	col	col
aldaketak	aldaketa	ABS	NUMP	**izan**	free	free
aldarrikapena	aldarrikapen	ABS	NUMS	**egin**	col	col
aldearekin	alde	SOZ	NUMS	**ekin**	free	free
aldeari	alde	DAT	NUMS	**eutsi**	free	free
alde	alde	ABS	MG	**jarri**	col	col

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
aldea	alde	ABS	NUMS	**kendu**	col	col
alde	alde	ABS	MG	**mintzatu**	col	col
alderdia	alderdi	ABS	NUMS	**utzi**	free	free
aldiz	aldi	INS	MG	**ari_izan**	free	free
alean	ale	INE	NUMS	**argitaratu**	free	free
alokairua	alokairu	ABS	NUMS	**ordaindu**	free	free
alternatiba	alternatiba	ABS	NUMS	**izan**	free	free
amak	ama	ABS	NUMP	**erauzi**	free	free
amaiera	amaiera	ABS	NUMS	**eman**	col	col
ametsa	amets	ABS	NUMS	**bete**	col	col
amua	amu	ABS	NUMS	**irentsi**	id	id
antxoa	antxoa	ABS	NUMS	**arrantzatu**	free	free
antxoa	antxoa	ABS	NUMS	**harrapatu**	free	free
antzerkiaz	antzerki	INS	NUMS	**gozatu**	free	free
antzezlana	antzezlan	ABS	NUMS	**taularatu**	col	col
apustuari	apustu	DAT	NUMS	**eutsi**	col	col
apustua	apustu	ABS	NUMS	**izan**	free	free
arantza	arantza	ABS	NUMS	**atera**	id	id
araudia	araudi	ABS	NUMS	**aldatu**	free	free
arauak	arau	ABS	NUMP	**hautsi**	col	col
arazoa	arazo	ABS	MG	**eduki**	free	free
arazoak	arazo	ABS	NUMP	**egon**	free	free
arazoak	arazo	ABS	NUMP	**eragin**	free	free
arazoa	arazo	ABS	NUMS	**konpondu**	free	free
arazoak	arazo	ABS	NUMP	**sortu**	free	free
ardura	ardura	ABS	NUMS	**hartu**	col	col
arduraz	ardura	INS	NUMS	**jokatu**	col	col
aretoa	areto	ABS	NUMS	**bete**	free	free
aretoan	areto	INE	NUMS	**paratu**	free	free
argazkian	argazki	INE	NUMS	**ikusi**	free	free
argazkiak	argazki	ABS	NUMP	**izan**	free	free
argibideak	argibide	ABS	NUMP	**eman**	col	col
argudioak	argudio	ABS	NUMP	**erabili**	col	col
armak	arma	ABS	NUMP	**entregatu**	free	free
armak	arma	ABS	NUMP	**saldu**	free	free
arnasa	arnasa	ABS	NUMS	**hartu**	col	col
arrakasta	arrakasta	ABS	NUMS	**ikusi**	free	free
arraun	arraun	ABS	MG	**egin**	col	id
arrazoiak	arrazoi	ABS	NUMP	**azaldu**	col	col
arretaz	arreta	INS	NUMS	**begiratu**	free	free
arriskuak	arrisku	ABS	NUMP	**hartu**	col	col
arriskuan	arrisku	INE	NUMS	**ikusi**	col	col
arriskuan	arrisku	INE	NUMS	**jarri**	col	col
arte	arte	ABS	MG	**atzeratu**	free	free
artean	arte	INE	NUMS	**aukeratu**	free	free
artean	arte	INE	NUMS	**banatu**	free	free
artean	arte	INE	NUMS	**hartu**	free	free

Bigrama normalizatua					**Sailkapena**	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
artetik	arte	ABL	NUMS	**igaro**	free	free
artelanak	artelan	ABS	NUMP	**erosi**	free	free
arte	arte	ABS	MG	**luzatu**	free	free
arte	arte	ABS	MG	**sufritu**	free	free
artikulu	artikulu	ABS	MG	**argitaratu**	free	free
artikuluan	artikulu	INE	NUMS	**oinarritu**	free	free
daskatasuna	askatasun	ABS	NUMS	**aldarrikatu**	free	free
askatasunez	askatasun	INS	NUMP	**erabaki**	free	free
askatasuna	askatasun	ABS	NUMS	**eskatu**	free	free
askatasuna	askatasun	ABS	NUMS	**exijitu**	free	free
askatasuna	askatasun	ABS	NUMS	**izan**	free	free
asmoa	asmo	ABS	NUMS	**agertu**	col	col
asmoa	asmo	ABS	NUMS	**izan**	col	col
asmoa	asmo	ABS	NUMS	**ukan**	col	col
asteburuan	asteburu	INE	NUMS	**jokatu**	free	free
aste	aste	ABS	MG	**eraman**	free	free
ataletan	atal	INE	NUMP	**banatu**	free	free
atentatua	atentatu	ABS	NUMS	**arbuiatu**	free	free
atentatua	atentatu	ABS	NUMS	**egin**	col	col
atentatua	atentatu	ABS	NUMS	**gertatu**	free	free
atsedenaldira	atsedenaldi	ALA	NUMS	**iritsi**	free	free
atxikimendua	atxikimendu	ABS	NUMS	**adierazi**	free	free
atxikimendua	atxikimendu	ABS	NUMS	**jaso**	col	col
atxiloketak	atxiloketa	ABS	NUMP	**salatu**	free	free
atzera	atze	ALA	NUMS	**begiratu**	id	id
atzera	atze	ALA	NUMS	**egin**	col	col
atzean	atze	INE	NUMS	**gelditu**	id	id
atzetik	atze	ABL	NUMS	**izan**	id	id
atzetik	atze	ABL	NUMS	**joan**	id	id
aukera	aukera	ABS	NUMS	**izan**	free	free
aulkian	aulki	INE	NUMS	**eseri**	free	free
aurka	aurka	ABS	MG	**ari_izan**	free	free
aurka	aurka	ABS	MG	**aritu**	free	free
aurka	aurka	ABS	MG	**azaldu**	col	col
aurka	aurka	ABS	MG	**borrokatu**	free	free
aurka	aurka	ABS	MG	**izan**	free	free
aurkezpena	aurkezpen	ABS	NUMS	**egin**	free	free
aurpegia	aurpegi	ABS	NUMS	**estali**	free	free
aurpegia	aurpegi	ABS	NUMS	**garbitu**	free	free
aurrean	aurre	INE	NUMS	**bildu**	free	free
aurrera	aurre	ALA	NUMS	**eraman**	id	id
aurrean	aurre	INE	NUMS	**ibili**	free	free
aurrera	aurre	ALA	NUMS	**jarraitu**	id	id
aurrean	aurre	INE	NUMS	**kikildu**	free	free
aurrelaria	aurrelari	ABS	NUMS	**fitxatu**	free	free
aurrelaria	aurrelari	ABS	NUMS	**izan**	free	free
autobusera	autobus	ALA	NUMS	**igo**	free	free

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
autobusetik	autobus	ABL	NUMS	**jaitsi**	free	free
autoa	auto	ABS	NUMS	**gidatu**	free	free
auto	auto	ABS	MG	**izan**	free	free
autoa	auto	ABS	NUMS	**lapurtu**	free	free
autopsia	autopsia	ABS	NUMS	**egin**	free	free
auzia	auzi	ABS	NUMS	**argitu**	free	free
auzia	auzi	ABS	NUMS	**artxibatu**	free	free
auzian	auzi	INE	NUMS	**auzipetu**	free	free
auzibidea	auzibidea	ABS	NUMS	**ireki**	col	col
auzitegiak	auzitegi	ABS	NUMP	**ebatzi**	free	free
auzitegira	auzitegi	ALA	NUMS	**jo**	free	free
auzian	auzi	INE	NUMS	**zigortu**	free	free
azpitik	azpi	ABL	NUMS	**egon**	id	id
babesa	babes	ABS	NUMS	**adierazi**	col	col
babesa	babes	ABS	NUMS	**izan**	free	free
baimena	baimen	ABS	NUMS	**izan**	free	free
bainua	bainu	ABS	NUMS	**hartu**	col	col
baitan	baita	INE	NUMS	**kokatu**	free	free
balantzea	balantze	ABS	NUMS	**egin**	col	col
baldintzak	baldintza	ABS	NUMP	**bete**	col	col
baldintzak	baldintza	ABS	NUMP	**hobetu**	free	free
balioa	balio	ABS	NUMS	**eman**	col	col
balioetan	balio	INE	NUMP	**oinarritu**	free	free
baloia	baloi	ABS	NUMS	**galdu**	col	col
baloia	baloi	ABS	NUMS	**kontrolatu**	free	free
baloia	baloi	ABS	NUMS	**lapurtu**	col	col
baloia	baloi	ABS	NUMS	**mugitu**	free	free
balorazioa	balorazio	ABS	NUMS	**egin**	col	col
barkamena	barkamen	ABS	NUMS	**eskatu**	col	col
barnean	barne	INE	NUMS	**egon**	free	free
barnean	barne	INE	NUMS	**kokatu**	free	free
barruan	barru	INE	NUMS	**egin**	free	free
batasunarekin	batasun	SOZ	NUMS	**bildu**	free	free
batasuna	batasun	ABS	NUMS	**lortu**	free	free
batzordea	batzorde	ABS	NUMS	**sortu**	free	free
bazkide	bazkide	ABS	MG	**izan**	col	col
bazterrak	bazter	ABS	NUMP	**harrotu**	id	id
bazterrak	bazter	ABS	NUMP	**nahastu**	id	id
bazterrean	bazter	INE	NUMS	**utzi**	id	id
begiak	begi	ABS	NUMP	**estali**	free	free
begiraleak	begirale	ABS	NUMP	**bidali**	free	free
beharra	behar	ABS	NUMS	**aldarrikatu**	free	free
beharra	behar	ABS	NUMS	**eduki**	col	col
beharra	behar	ABS	NUMS	**ikusi**	free	free
beharra	behar	ABS	NUMS	**nabarmendu**	free	free
behar	behar	ABS	MG	**ukan**	col	id
behera	behe	ALA	NUMS	**erori**	col	col

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
beldurra	beldur	ABS	NUMS	**eragin**	col	col
berdinketa	berdinketa	ABS	NUMS	**lortu**	col	col
berriari	berri	DAT	NUMS	**adierazi**	free	free
berri	berri	ABS	MG	**izan**	free	free
bertsoak	bertso	ABS	NUMP	**kantatu**	free	free
besotik	beso	ABL	NUMS	**heldu**	free	free
betoa	beto	ABS	NUMS	**jarri**	col	col
bidaia	bidaia	ABS	NUMS	**izan**	free	free
bidetik	bide	ABL	NUMS	**atera**	id	id
bideari	bide	DAT	NUMS	**ekin**	id	id
bidea	bide	ABS	NUMS	**eman**	id	id
bidea	bide	ABS	NUMS	**hasi**	col	col
bidetik	bide	ABL	NUMS	**joan**	col	col
bidesaria	bidesari	ABS	NUMS	**kendu**	free	free
bihotzekoak	bihotzeko	ABS	NUMP	**jo**	col	col
bikotekidea	bikotekide	ABS	NUMS	**jo**	free	free
biktima	biktima	ABS	MG	**eragin**	col	col
biktima	biktima	ABS	NUMS	**izan**	free	free
bileran	bilera	INE	NUMS	**izan**	free	free
biran	bira	INE	NUMS	**ibili**	free	free
birusaz	birus	INS	NUMS	**kutsatu**	free	free
bisita	bisita	ABS	NUMS	**gidatu**	free	free
bisita	bisita	ABS	MG	**izan**	free	free
bistan	bista	INE	NUMS	**izan**	col	col
bizi-baldintzak	bizi-baldintza	ABS	NUMP	**hobetu**	free	free
bizi-bizirik	bizi-bizi	ABS	MG	**egon**	free	free
bizia	bizi	ABS	NUMS	**galdu**	col	col
bizi	bizi	ABS	MG	**izan**	col	id
bizikletan	bizikleta	INE	NUMS	**ibili**	free	free
bizipenak	bizipen	ABS	NUMP	**kontatu**	free	free
bizitza	bizitza	ABS	NUMS	**aldatu**	free	free
bizitza	bizitza	ABS	NUMS	**kontatu**	free	free
bizitza	bizitza	ABS	NUMS	**salbatu**	free	free
bizkarrezurrari	bizkarrezur	DAT	NUMS	**eutsi**	free	free
biztanle	biztanle	ABS	MG	**ukan**	free	free
borondatea	borondate	ABS	NUMS	**erakutsi**	col	col
borondatea	borondate	ABS	NUMS	**izan**	free	free
borondatea	borondate	ABS	NUMS	**ukan**	free	free
borrokan	borroka	INE	NUMS	**aritu**	free	free
borrokan	borroka	INE	NUMS	**jarraitu**	free	free
borrokan	borroka	INE	NUMS	**sartu**	col	col
boterean	botere	INE	NUMS	**egon**	col	col
boterera	botere	ALA	NUMS	**heldu**	free	free
botereaz	botere	INS	NUMS	**jabetu**	free	free
boteretik	botere	ABL	NUMS	**kendu**	free	free
boterea	botere	ABS	NUMS	**partekatu**	free	free
botoa	boto	ABS	NUMS	**eskatu**	free	free

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
botoekin	boto	SOZ	NUMP	**onartu**	free	free
boza	boz	ABS	NUMS	**eman**	col	col
bozeramaileak	bozeramaile	ERG	NUMS	**adierazi**	free	free
bozeramaileak	bozeramaile	ERG	NUMS	**jakinarazi**	free	free
bozetan	boz	INE	NUMP	**nagusitu**	free	free
bueltan	buelta	INE	NUMS	**ibili**	id	id
buru-belarri	buru-belarri	ABS	MG	**murgildu**	free	free
buruarekin	buru	SOZ	NUMS	**bildu**	free	free
buruan	buru	INE	NUMS	**egon**	id	id
burua	buru	ABS	NUMS	**estali**	free	free
buruan	buru	INE	NUMS	**jarri**	free	free
burutik	buru	ABL	NUMS	**kendu**	id	id
burua	buru	ABS	NUMS	**makurtu**	id	id
burua	buru	ABS	NUMS	**zuritu**	id	id
dantza	dantza	ABS	NUMS	**izan**	free	free
dantzan	dantza	INE	NUMS	**jarri**	id	id
datuei	datu	DAT	NUMP	**begiratu**	free	free
datuak	datu	ABS	NUMP	**eman**	free	free
datuak	datu	ABS	NUMP	**izan**	free	free
debekua	debeku	ABS	NUMS	**ezarri**	col	col
debekua	debeku	ABS	NUMS	**kendu**	free	free
defentsan	defentsa	INE	NUMS	**aritu**	free	free
defentsa	defentsa	ABS	NUMS	**gogortu**	free	free
defentsan	defentsa	INE	NUMS	**oinarritu**	free	free
defizita	defizit	ABS	NUMS	**murriztu**	free	free
deialdia	deialdi	ABS	NUMS	**egin**	col	col
deialdian	deialdi	INE	NUMS	**sartu**	col	col
deiari	dei	DAT	NUMS	**erantzun**	free	free
deklarazioetan	deklarazio	INE	NUMP	**oinarritu**	free	free
dekretua	dekretu	ABS	NUMS	**baliogabetu**	free	free
delitua	delitu	ABS	NUMS	**egin**	col	col
delitua	delitu	ABS	NUMS	**izan**	free	free
delitutzat	delitu	PRO	MG	**jo**	free	free
delitua	delitu	ABS	NUMS	**leporatu**	free	free
denboraldia	denboraldi	ABS	NUMS	**bukatu**	free	free
denboraldia	denboraldi	ABS	NUMS	**izan**	free	free
denboraldia	denboraldi	ABS	NUMS	**ukan**	free	free
dimisioa	dimisio	ABS	NUMS	**aurkeztu**	col	col
dimisioa	dimisio	ABS	NUMS	**eman**	col	col
diskoa	disko	ABS	NUMS	**atera**	col	col
diskoa	disko	ABS	NUMS	**grabatu**	free	free
doktrina	doktrina	ABS	MG	**aplikatu**	free	free
dolar	dolar	ABS	MG	**ordaindu**	free	free
dolarretan	dolar	INE	NUMP	**saldu**	free	free
domina	domina	ABS	MG	**lortu**	free	free
dudak	duda	ABS	NUMP	**argitu**	free	free
ebazpena	ebazpen	ABS	NUMS	**onartu**	free	free

Bigrama normalizatua					**Sailkapena**	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
efektua	efektu	ABS	NUMS	**eragin**	free	free
egoeratik	egoera	ABL	NUMS	**atera**	free	free
egoera	egoera	ABS	NUMS	**azaldu**	free	free
egoera	egoera	ABS	NUMS	**aztertu**	free	free
egoerara	egoera	ALA	NUMS	**egokitu**	free	free
egoeran	egoera	INE	NUMS	**egon**	free	free
egoera	egoera	ABS	NUMS	**gainditu**	col	col
egoera	egoera	ABS	NUMS	**hobetu**	free	free
egoerara	egoera	ALA	NUMS	**itzuli**	free	free
egoeraz	egoera	INS	NUMS	**kezkatu**	free	free
egoera	egoera	ABS	NUMS	**legeztatu**	free	free
egoerak	egoera	ABS	NUMP	**okerreratu**	free	free
egoera	egoera	ABS	NUMS	**okertu**	free	free
egoitzan	egoitza	INE	NUMS	**egin**	free	free
egokia	egoki	ABS	NUMS	**iruditu**	free	free
eguerdian	eguerdi	INE	NUMS	**elkarretaratu**	free	free
eguna	egun	ABS	NUMS	**izan**	free	free
egunkariari	egunkari	DAT	NUMS	**adierazi**	free	free
egunkariak	egunkari	ABS	NUMP	**argitaratu**	free	free
egunkaria	egunkari	ABS	NUMS	**irakurri**	free	free
egunkarian	egunkari	INE	NUMS	**irakurri**	free	free
egurra	egur	ABS	NUMS	**eman**	id	id
egutegia	egutegi	ABS	NUMS	**zehaztu**	free	free
ekintzak	ekintza	ABS	NUMP	**egin**	free	free
ekintzetara	ekintza	ALA	NUMP	**igaro**	col	col
ekintzetara	ekintza	ALA	NUMP	**pasatu**	col	col
ekitaldi	ekitaldi	ABS	MG	**antolatu**	free	free
ekitaldi	ekitaldi	ABS	MG	**prestatu**	free	free
ekoizpena	ekoizpen	ABS	NUMS	**handitu**	free	free
ekonomia	ekonomia	ABS	NUMS	**suspertu**	free	free
elebitasuna	elebitasun	ABS	NUMS	**bermatu**	free	free
eledunak	eledun	ABS	NUMP	**zehaztu**	free	free
elementua	elementu	ABS	NUMS	**izan**	free	free
elizatik	eliza	ABL	NUMS	**atera**	free	free
elkarlanean	elkarlan	INE	NUMS	**aritu**	free	free
elkarlanean	elkarlan	INE	NUMS	**egin**	free	free
elkarrizketan	elkarrizketa	INE	NUMS	**eskaini**	free	free
elkartasuna	elkartasun	ABS	NUMS	**adierazi**	col	col
elkarteak	elkarte	ABS	NUMP	**antolatu**	free	free
emaitzei	emaitza	DAT	NUMP	**begiratu**	free	free
emaitzek	emaitza	ERG	NUMP	**lagundu**	free	free
emaitzetan	emaitza	INE	NUMP	**oinarritu**	free	free
emakumeak	emakume	ABS	NUMP	**izan**	free	free
emakumeek	emakume	ERG	NUMP	**jasan**	free	free
emakumeek	emakume	ERG	NUMP	**pairatu**	free	free
emanaldia	emanaldi	ABS	NUMS	**egin**	col	col
emaztea	emazte	ABS	NUMS	**izan**	free	free

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
enbor	enbor	ABS	MG	**ebaki**	free	free
enkantean	enkante	INE	NUMS	**jarri**	col	col
enpleguari	enplegu	DAT	NUMS	**eutsi**	free	free
entrenamendua	entrenamendu	ABS	NUMS	**egin**	col	col
epaiketa	epaiketa	ABS	NUMS	**hasi**	free	free
epaimahaiak	epaimahai	ABS	NUMP	**osatu**	free	free
epaitegietara	epaitegi	ALA	NUMP	**jo**	free	free
epea	epe	ABS	NUMS	**agortu**	col	col
epea	epe	ABS	NUMS	**jarri**	free	free
erabilera	erabilera	ABS	NUMS	**sustatu**	free	free
eraikina	eraikin	ABS	NUMS	**hustu**	free	free
erakusketa	erakusketa	ABS	NUMS	**inauguratu**	free	free
erakusketa	erakusketa	ABS	NUMS	**izan**	free	free
erakusketa	erakusketa	ABS	NUMS	**zabaldu**	free	free
erakustaldia	erakustaldi	ABS	NUMS	**egin**	col	col
erakustaldia	erakustaldi	ABS	NUMS	**eman**	col	col
erantzukizunak	erantzukizun	ABS	NUMP	**eskatu**	free	free
erantzuletzat	erantzule	PRO	MG	**jo**	free	free
erantzuna	erantzun	ABS	NUMS	**eman**	col	col
erantzuna	erantzun	ABS	NUMS	**jaso**	col	col
erasoari	eraso	DAT	NUMS	**ekin**	free	free
erasoa	eraso	ABS	NUMS	**gaitzetsi**	free	free
erasoa	eraso	ABS	NUMS	**gertatu**	free	free
erasoa	eraso	ABS	NUMS	**jo**	col	col
erasotzat	eraso	PRO	MG	**jo**	free	free
erasotzailea	erasotzaile	ABS	NUMS	**atxilotu**	free	free
erasoak	eraso	ABS	NUMP	**ugaritu**	free	free
erdira	erdi	ALA	NUMS	**jaitsi**	free	free
erditik	erdi	ABL	NUMS	**kendu**	id	id
erdira	erdi	ALA	NUMS	**murriztu**	free	free
eredua	eredu	ABS	NUMS	**inposatu**	free	free
erosketak	erosketa	ABS	NUMP	**egin**	col	col
erraztasunak	erraztasun	ABS	NUMP	**eman**	col	col
errealitatea	errealitate	ABS	NUMS	**islatu**	free	free
erreferenduma	erreferendum	ABS	NUMS	**egin**	free	free
erreferentzia	erreferentzia	ABS	NUMS	**bilakatu**	free	free
erreferentzia	erreferentzia	ABS	NUMS	**izan**	col	col
errege	errege	ABS	MG	**izendatu**	free	free
errepidetik	errepide	ABL	NUMS	**atera**	free	free
errepidea	errepide	ABS	NUMS	**itxi**	free	free
errepidean	errepide	INE	NUMS	**jazo**	free	free
errepidea	errepide	ABS	NUMS	**zeharkatu**	free	free
errepresioa	errepresio	ABS	NUMS	**areagotu**	free	free
errotik	erro	ABL	NUMS	**aldatu**	id	id
erroetara	erro	ALA	NUMP	**jo**	id	id
errotik	erro	ABL	NUMS	**moztu**	id	id
eskaera	eskaera	ABS	NUMS	**egin**	col	col

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
eskaera	eskaera	ABS	NUMS	**izan**	free	free
eskaintza	eskaintza	ABS	NUMS	**onartu**	free	free
eskasa	eskas	ABS	NUMS	**izan**	free	free
esker	esker	ABS	MG	**lortu**	free	free
eskolatik	eskola	ABL	NUMS	**atera**	free	free
eskola	eskola	ABS	NUMS	**izan**	free	free
eskubidea	eskubide	ABS	NUMS	**aldarrikatu**	free	free
eskubidea	eskubide	ABS	NUMS	**erabili**	free	free
eskubidean	eskubide	INE	NUMS	**oinarritu**	free	free
esku	esku	ABS	MG	**egon**	free	free
eskutik	esku	ABL	NUMS	**etorri**	id	id
eskuetan	esku	INE	NUMP	**hartu**	id	id
eskutik	esku	ABL	NUMS	**heldu**	id	id
esku	esku	ABS	MG	**izan**	free	free
eskuetan	esku	INE	NUMP	**izan**	id	id
eskua	esku	ABS	NUMS	**luzatu**	id	id
eskua	esku	ABS	NUMS	**sartu**	id	id
esku	esku	ABS	MG	**ukan**	id	id
eskuetan	esku	INE	NUMP	**ukan**	id	id
espedientea	espediente	ABS	NUMS	**aurkeztu**	col	col
esperientziak	esperientzia	ABS	NUMP	**partekatu**	free	free
espero	espero	ABS	MG	**izan**	col	id
espiritua	espiritu	ABS	NUMS	**berreskura**	free	free
estatutua	estatutu	ABS	NUMS	**bete**	free	free
estatutua	estatutu	ABS	NUMS	**erreformatu**	free	free
estradizioa	estradizio	ABS	NUMS	**eskatu**	free	free
etapa	etapa	ABS	MG	**irabazi**	free	free
etapa	etapa	ABS	MG	**jokatu**	free	free
etorkin	etorkin	ABS	MG	**atzeman**	free	free
etorkizunari	etorkizun	DAT	NUMS	**begiratu**	col	col
etorkizunaz	etorkizun	INS	NUMS	**erabaki**	free	free
etorkizuna	etorkizun	ABS	NUMS	**eraiki**	id	id
etorkizuna	etorkizun	ABS	NUMS	**hipotekatu**	free	free
etorkizuna	etorkizun	ABS	NUMS	**izan**	free	free
etxetik	etxe	ABL	NUMS	**atera**	free	free
etxean	etxe	INE	NUMS	**atxiki**	free	free
etxean	etxe	INE	NUMS	**atxilotu**	free	free
etxebizitza	etxebizitza	ABS	MG	**miatu**	free	free
etxera	etxe	ALA	NUMS	**bueltatu**	free	free
etxean	etxe	INE	NUMS	**eduki**	free	free
etxera	etxe	ALA	NUMS	**eraman**	free	free
etxeari	etxe	DAT	NUMS	**eraso**	free	free
etxean	etxe	INE	NUMS	**galdu**	free	free
etxetik	etxe	ABL	NUMS	**irten**	free	free
etxera	etxe	ALA	NUMS	**joan**	free	free
etxea	etxe	ABS	NUMS	**miatu**	free	free
etxe	etxe	ABS	MG	**suntsitu**	free	free

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
etxea	etxe	ABS	NUMS	**utzi**	free	free
eurora	euro	ALA	NUMS	**egokitu**	free	free
euro	euro	ABS	MG	**eman**	free	free
euro	euro	ABS	MG	**fakturatu**	free	free
euskaraz	euskara	INS	NUMS	**aritu**	free	free
euskaraz	euskara	INS	NUMS	**bizitu**	free	free
euskarari	euskara	DAT	NUMS	**egon**	free	free
euskarara	euskara	ALA	NUMS	**hurbildu**	free	free
euskarara	euskara	ALA	NUMS	**itzuli**	free	free
ezaguna	ezagun	ABS	NUMS	**izan**	col	col
ezaugarrietara	ezaugarri	ALA	NUMP	**egokitu**	free	free
ezaugarriak	ezaugarri	ABS	NUMP	**ukan**	free	free
ezbeharra	ezbehar	ABS	NUMS	**gertatu**	free	free
ezean	ez	INE	NUMS	**ibili**	free	free
ezinbestekoa	ezinbeste	GELABS	NUMS	**izan**	free	free
ezinbestekotzat	ezinbeste	GELPRO	MG	**jo**	col	col
ezin	ezin	ABS	MG	**esan**	free	free
ezin	ezin	ABS	MG	**izan**	col	id
ezkutuan	ezkutu	INE	NUMS	**gorde**	free	free
eztabaida	eztabaida	ABS	NUMS	**egin**	col	col
eztabaida	eztabaida	ABS	NUMS	**izan**	col	col
eztabaida	eztabaida	ABS	NUMS	**sortu**	col	col
ezustean	ezuste	INE	NUMS	**harrapatu**	id	id
ezustekoa	ezusteko	ABS	MG	**izan**	free	free
falta	falta	ABS	MG	**egin**	col	col
falta	falta	ABS	MG	**izan**	col	id
festan	festa	INE	NUMS	**murgildu**	col	col
filosofia	filosofia	ABS	NUMS	**ikasi**	free	free
finalerdietaraino	finalerdi	ABU	NUMP	**iritsi**	free	free
finalerdietara	finalerdi	ALA	NUMP	**iritsi**	free	free
finalerdietarako	finalerdi	ALAGEL	NUMP	**sailkatu**	free	free
finalera	final	ALA	NUMS	**heldu**	free	free
finala	final	ABS	NUMS	**izan**	free	free
final-laurdenetara	final-laurden	ALA	NUMP	**sailkatu**	free	free
final-zortzirenetara	final-zortziren	ALA	NUMP	**sailkatu**	free	free
frantsesez	frantses	INS	NUMP	**idatzi**	free	free
froga	froga	ABS	NUMS	**izan**	free	free
funtzioa	funtzio	ABS	NUMS	**bete**	col	col
futbolean	futbol	INE	NUMS	**jokatu**	free	free
gabe	gabe	ABS	MG	**gelditu**	free	free
gaiaz	gai	INS	NUMS	**mintzatu**	free	free
gainbehera	gainbehera	ABS	NUMS	**etorri**	col	col
gaindi	gaindi	ABS	MG	**ibili**	free	free
gainean	gain	INE	NUMS	**egon**	id	id
gain	gain	ABS	MG	**hartu**	id	id
gainetik	gain	ABL	NUMS	**kendu**	id	id
gainetik	gain	ABL	NUMS	**pasatu**	id	id

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
gaitasuna	gaitasun	ABS	NUMS	**ukan**	free	free
gaitza	gaitz	ABS	NUMS	**sendatu**	free	free
gaixotasunak	gaixotasun	ABS	NUMP	**sendatu**	free	free
galbahetik	galbahe	ABL	NUMS	**pasatu**	id	id
galderak	galdera	ABS	NUMP	**pausatu**	col	col
garaia	garai	ABS	NUMS	**iritsi**	free	free
garaipena	garaipen	ABS	NUMS	**eskuratu**	col	col
garaipena	garaipen	ABS	NUMS	**lortu**	col	col
gardentasuna	gardentasun	ABS	NUMS	**bermatu**	free	free
garrantzia	garrantzi	ABS	NUMS	**eduki**	free	free
garrantzia	garrantzi	ABS	NUMS	**eman**	col	col
garrantzia	garrantzi	ABS	NUMS	**kendu**	col	col
garrantziaz	garrantzi	INS	NUMS	**ohartarazi**	free	free
garrantzia	garrantzi	ABS	NUMS	**ukan**	free	free
gasak	gasa	ABS	NUMP	**murriztu**	free	free
gatazka	gatazka	ABS	NUMS	**konpondu**	free	free
gehiengoa	gehiengo	ABS	NUMS	**izan**	free	free
gehiengoa	gehiengo	ABS	NUMS	**lortu**	col	col
gelatik	gela	ABL	NUMS	**irten**	free	free
geldialdia	geldialdi	ABS	NUMS	**egin**	free	free
genozidioa	genozidio	ABS	NUMS	**leporatu**	free	free
gerra	gerra	ABS	NUMS	**amaitu**	free	free
gerrikoa	gerriko	ABS	NUMS	**estutu**	id	id
gezurra	gezur	ABS	NUMS	**esan**	col	col
gezurtzat	gezur	PRO	MG	**jo**	free	free
gidaria	gidari	ABS	NUMS	**izan**	free	free
giltza	giltza	ABS	NUMS	**izan**	id	id
giroa	giro	ABS	NUMS	**berotu**	id	id
giroa	giro	ABS	NUMS	**sortu**	col	col
gisa	gisa	ABS	NUMS	**deskribatu**	free	free
gisa	gisa	ABS	MG	**egin**	free	free
gisa	gisa	ABS	MG	**izan**	free	free
gitarra	gitarra	ABS	NUMS	**jo**	free	free
gizartea	gizarte	ABS	NUMS	**mobilizatu**	free	free
gizartea	gizarte	ABS	NUMS	**sentsibilizatu**	free	free
gizarteak	gizarte	ABS	NUMP	**ukan**	free	free
gizartea	gizarte	ABS	NUMS	**zatitu**	free	free
gizon	gizon	ABS	MG	**atxilotu**	free	free
gizonezko	gizonezko	ABS	MG	**atxilotu**	free	free
gizonez	gizon	INS	NUMP	**jantzi**	free	free
gobernuari	gobernu	DAT	NUMS	**eskatu**	free	free
gobernura	gobernu	ALA	NUMS	**itzuli**	free	free
gobernua	gobernu	ABS	NUMS	**izan**	free	free
gobernuan	gobernu	INE	NUMS	**izan**	free	free
gobernuak	gobernu	ABS	NUMP	**jakinarazi**	free	free
gobernua	gobernu	ABS	NUMS	**osatu**	free	free
gobernua	gobernu	ABS	NUMS	**sostengatu**	col	col

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
gogoa	gogo	ABS	NUMS	**eduki**	col	col
gogoan	gogo	INE	NUMS	**hartu**	col	col
gogoa	gogo	ABS	NUMS	**ukan**	col	col
gogoan	gogo	INE	NUMS	**ukan**	col	col
goizean	goiz	INE	NUMS	**egin**	free	free
gola	gol	ABS	MG	**jaso**	col	col
gol	gol	ABS	MG	**sartu**	col	col
gorabeherak	gorabehera	ABS	NUMP	**izan**	free	free
gorriak	gorri	ABS	NUMP	**ikusi**	id	id
granadak	granada	ABS	NUMP	**jaurti**	free	free
grebara	greba	ALA	NUMS	**deitu**	free	free
greban	greba	INE	NUMS	**egon**	col	col
grebari	greba	DAT	MG	**ekin**	free	free
greba	greba	ABS	MG	**hasi**	free	free
gripeaz	gripe	INS	NUMS	**kutsatu**	free	free
gunea	gune	ABS	NUMS	**izan**	free	free
gustura	gustu	ALA	NUMS	**agertu**	col	col
gutxiengoan	gutxiengo	INE	NUMS	**gobernatu**	free	free
hamarkadan	hamarkada	INE	NUMS	**hasi**	free	free
hankapean	hankape	INE	NUMS	**ibili**	id	id
hanka	hanka	ABS	NUMS	**sartu**	id	id
haratago	haratago	ABS	MG	**joan**	id	id
harkaitzean	harkaitz	INE	NUMS	**eskalatu**	free	free
harremanak	harreman	ABS	NUMP	**eten**	col	col
harremanak	harreman	ABS	NUMP	**gaiztotu**	col	col
harremanak	harreman	ABS	NUMP	**haustu**	col	col
harremanak	harreman	ABS	NUMP	**izan**	free	free
harremanak	harreman	ABS	NUMP	**landu**	free	free
harresia	harresi	ABS	NUMS	**eraiki**	id	id
harresia	harresi	ABS	NUMS	**erori**	id	id
harridura	harridura	ABS	NUMS	**sortu**	col	col
harrobian	harrobi	INE	NUMS	**hazi**	id	id
hasieratik	hasiera	ABL	NUMS	**aritu**	free	free
hasieran	hasiera	INE	NUMS	**aurreikusi**	free	free
hasiera	hasiera	ABS	NUMS	**orekatu**	free	free
hastapenetik	hastapen	ABL	NUMS	**ari_izan**	free	free
hausnarketa	hausnarketa	ABS	NUMS	**egin**	col	col
hauteskundeetan	hauteskunde	INE	NUMP	**aurkeztu**	free	free
hauteskundeak	hauteskunde	ABS	NUMP	**izan**	free	free
hautsak	hauts	ABS	NUMP	**harrotu**	id	id
hegalean	hegal	INE	NUMS	**jokatu**	free	free
hegazkinetik	hegazkin	ABL	NUMS	**jaitsi**	free	free
hegoaldean	hegoalde	INE	NUMS	**egon**	free	free
hektarea	hektarea	ABS	MG	**erre**	free	free
helbidera	helbide	ALA	NUMS	**idatzi**	free	free
helburua	helburu	ABS	NUMS	**lortu**	free	free
helduak	heldu	ABS	NUMP	**euskaldundu**	free	free

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
helegitea	helegite	ABS	NUMS	**aurkeztu**	col	col
helegitea	helegite	ABS	NUMS	**aztertu**	free	free
helegitea	helegite	ABS	NUMS	**jarri**	col	col
heriotzara	heriotza	ALA	NUMS	**kondenatu**	col	col
herrialdeekin	herrialde	SOZ	NUMP	**alderatu**	free	free
herrialdeetara	herrialde	ALA	NUMP	**bidali**	free	free
herrialdea	herrialde	ABS	NUMS	**okupatu**	free	free
herrian	herri	INE	NUMS	**bizitu**	free	free
herria	herri	ABS	NUMS	**eraiki**	id	id
herriari	herri	DAT	NUMS	**eraso**	free	free
herria	herri	ABS	NUMS	**ezagutu**	free	free
herrira	herri	ALA	NUMS	**iritsi**	free	free
herrira	herri	ALA	NUMS	**itzuli**	free	free
herria	herri	ABS	NUMS	**izan**	free	free
herrira	herri	ALA	NUMS	**joan**	free	free
herritarrak	herritar	ABS	NUMP	**beldurtu**	free	free
herritarrak	herritar	ABS	NUMP	**informatu**	free	free
herritarrak	herritar	ABS	NUMP	**izan**	free	free
hilabete	hilabete	ABS	MG	**eraman**	free	free
hilabetez	hilabete	INS	MG	**luzatu**	free	free
hipoteka	hipoteka	ABS	NUMS	**ordaindu**	free	free
hiria	hiri	ABS	NUMS	**bonbardatu**	free	free
hirian	hiri	INE	NUMS	**egin**	free	free
historia	historia	ABS	NUMS	**errepikatu**	col	col
hitzaldiak	hitzaldi	ABS	NUMP	**antolatu**	free	free
hitzaldia	hitzaldi	ABS	NUMS	**eman**	col	col
hitz	hitz	ABS	MG	**aritu**	free	free
hitzarmena	hitzarmen	ABS	NUMS	**sinatu**	free	free
hitzak	hitz	ABS	NUMP	**esan**	free	free
hondoa	hondo	ABS	NUMS	**jo**	id	id
hortzak	hortz	ABS	NUMP	**erakutsi**	id	id
hortzak	hortz	ABS	NUMP	**estutu**	id	id
hotsak	hots	ABS	NUMP	**entzun**	free	free
ibaira	ibai	ALA	NUMS	**erori**	free	free
ibilbidea	ibilbide	ABS	NUMS	**izan**	free	free
ibilbidea	ibilbide	ABS	NUMS	**osatu**	free	free
ibilbidea	ibilbide	ABS	NUMS	**saritu**	free	free
idazlea	idazle	ABS	NUMS	**izan**	free	free
idazlea	idazle	ABS	NUMS	**jaio**	free	free
ideiak	ideia	ABS	NUMP	**defendatu**	free	free
igandean	igande	INE	NUMS	**izan**	free	free
igandean	igande	INE	NUMS	**jokatu**	free	free
igoera	igoera	ABS	NUMS	**izan**	free	free
ikasketak	ikasketak	ABS	MG	**egin**	col	col
ikasleak	ikasle	ABS	NUMP	**izan**	free	free
ikastaroak	ikastaro	ABS	NUMP	**eskaini**	free	free
ikerketak	ikerketa	ABS	NUMP	**egin**	free	free

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
ikur	ikur	ABS	MG	**bihurtu**	col	col
ikur	ikur	ABS	MG	**bilakatu**	col	col
ikuskizuna	ikuskizun	ABS	NUMS	**eskaini**	col	col
ikuslea	ikusle	ABS	NUMS	**harrapatu**	id	id
ilea	ile	ABS	NUMS	**moztu**	free	free
indarrak	indar	ABS	NUMP	**ahitu**	free	free
indarrak	indar	ABS	NUMP	**bildu**	col	col
indarrean	indar	INE	NUMS	**egon**	col	col
indarra	indar	ABS	NUMS	**galdu**	free	free
indarkeria	indarkeria	ABS	NUMS	**arbuiatu**	free	free
indarkeriarekin	indarkeria	SOZ	MG	**lotu**	free	free
indarrak	indar	ABS	NUMP	**neurtu**	id	id
indarrez	indar	INS	NUMP	**sakabanatu**	free	free
indarra	indar	ABS	NUMS	**ukan**	free	free
independentzia	independentzia	ABS	NUMS	**aldarrikatu**	free	free
informazioa	informazio	ABS	NUMS	**ezkutatu**	free	free
informazioa	informazio	ABS	NUMS	**trukatu**	free	free
ingelesez	ingeles	INS	NUMP	**idatzi**	free	free
ingelesa	ingeles	ABS	NUMS	**ikasi**	free	free
ingelesa	ingeles	ABS	NUMS	**irakatsi**	free	free
ingelesera	ingeles	ALA	NUMS	**itzuli**	free	free
inguruan	inguru	INE	NUMS	**aritu**	free	free
inguruan	inguru	INE	NUMS	**bildu**	free	free
inguruan	inguru	INE	NUMS	**ibili**	free	free
inguru	inguru	ABS	MG	**izan**	free	free
ingurukoa	inguru	GELABS	NUMS	**izan**	free	free
inguruan	inguru	INE	NUMS	**mintzatu**	free	free
inkesta	inkesta	ABS	NUMS	**egin**	col	col
integrazioa	integrazio	ABS	NUMS	**bermatu**	free	free
interesak	interes	ABS	NUMP	**defenditu**	free	free
interesak	interes	ABS	NUMP	**lehenetsi**	free	free
interesa	interes	ABS	NUMS	**ukan**	col	col
iraunkortasuna	iraunkortasun	ABS	NUMS	**bermatu**	free	free
iraupena	iraupen	ABS	NUMS	**izan**	free	free
iritzia	iritzi	ABS	NUMS	**aldatu**	free	free
iritzia	iritzi	ABS	NUMS	**azaldu**	free	free
iritzia	iritzi	ABS	NUMS	**entzun**	free	free
iritzia	iritzi	ABS	NUMS	**jaso**	free	free
iritzia	iritzi	ABS	NUMS	**plazaratu**	free	free
ironia	ironia	ABS	NUMS	**erabili**	free	free
irudia	irudi	ABS	NUMS	**eman**	col	col
irudia	irudi	ABS	NUMS	**izan**	free	free
irudimena	irudimen	ABS	NUMS	**landu**	free	free
irudia	irudi	ABS	NUMS	**zikindu**	id	id
iruzurra	iruzur	ABS	NUMS	**egin**	col	col
isiltasuna	isiltasun	ABS	NUMS	**nagusitu**	free	free
ispiluan	ispilu	INE	NUMS	**begiratu**	free	free

Bigrama normalizatua					**Sailkapena**	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
istripuz	istripu	INS	MG	**hil**	free	free
isuna	isun	ABS	NUMS	**jarri**	col	col
isurketak	isurketa	ABS	NUMP	**gutxitu**	free	free
ituna	itun	ABS	NUMS	**izenpetu**	free	free
ituna	itun	ABS	NUMS	**sinatu**	free	free
itxaropena	itxaropen	ABS	NUMS	**galdu**	col	col
itxialdia	itxialdi	ABS	NUMS	**egin**	free	free
itxura	itxura	ABS	NUMS	**eman**	col	col
itzulian	itzuli	INE	NUMS	**lehiatu**	free	free
izenarekin	izen	SOZ	NUMS	**ezagutu**	free	free
jaialdia	jaialdi	ABS	NUMS	**antolatu**	free	free
jaiez	jai	INS	NUMP	**gozatu**	free	free
jakin-mina	jakin-min	ABS	NUMS	**ase**	col	col
jakin-mina	jakin-min	ABS	NUMS	**piztu**	col	col
jaramon	jaramon	ABS	MG	**egin**	col	id
jardunaldiak	jardunaldi	ABS	NUMP	**egin**	free	free
jarduna	jardun	ABS	NUMS	**eten**	free	free
jarraipena	jarraipen	ABS	NUMS	**izan**	free	free
jarrera	jarrera	ABS	NUMS	**agertu**	free	free
jarrera	jarrera	ABS	NUMS	**argitu**	free	free
jarrera	jarrera	ABS	NUMS	**azaldu**	free	free
jarrera	jarrera	ABS	NUMS	**erakutsi**	col	col
jarrera	jarrera	ABS	NUMS	**gaitzetsi**	col	col
jarrera	jarrera	ABS	NUMS	**kritikatu**	free	free
jarrera	jarrera	ABS	NUMS	**salatu**	free	free
jasoaldi	jasoaldi	ABS	MG	**egin**	col	col
jaurlaritzari	jaurlaritza	DAT	MG	**eskatu**	free	free
jaurtiketa	jaurtiketa	ABS	MG	**gauzatu**	col	col
jazarpena	jazarpen	ABS	NUMS	**pairatu**	free	free
jazarpena	jazarpen	ABS	NUMS	**salatu**	free	free
jendea	jende	ABS	NUMS	**animatu**	free	free
jendea	jende	ABS	NUMS	**beldurtu**	free	free
jendearengana	jende	ALA	NUMS	**heldu**	col	col
jendea	jende	ABS	NUMS	**hunkitu**	free	free
jendez	jende	INS	MG	**inguratu**	free	free
jendea	jende	ABS	NUMS	**nazkatu**	free	free
jendetza	jendetza	ABS	NUMS	**bildu**	free	free
joan-etorrian	joan-etorri	INE	NUMS	**ibili**	col	col
joera	joera	ABS	NUMS	**izan**	col	col
joera	joera	ABS	MG	**ukan**	col	col
jokaera	jokaera	ABS	NUMS	**salatu**	free	free
jokalariei	jokalari	DAT	NUMP	**egon**	free	free
jokoan	joko	INE	NUMS	**izan**	col	col
jokoan	joko	INE	NUMS	**sartu**	col	col
kalera	kale	ALA	NUMS	**atera**	id	id
kalean	kale	INE	NUMS	**ibili**	free	free
kalejira	kalejira	ABS	NUMS	**egin**	col	col

Bigrama normalizatua					Sailkapena	
izena				aditza	(1)	(2)
forma	lema	kasua	mugat.			
kalitatea	kalitate	ABS	NUMS	**bermatu**	col	col
kalitatea	kalitate	ABS	NUMS	**hobetu**	free	free
kalte	kalte	ABS	MG	**egin**	col	id
kalteak	kalte	ABS	NUMP	**eragin**	col	col
kalteak	kalte	ABS	NUMP	**konpondu**	free	free
kanpainan	kanpaina	INE	NUMS	**agindu**	free	free
kanpainari	kanpaina	DAT	MG	**ekin**	free	free
kanpotik	kanpo	ABL	NUMS	**ekarri**	free	free
kanpoan	kanpo	INE	NUMS	**izan**	free	free
kanpokotzat	kanpoko	PRO	MG	**jo**	free	free
kantak	kanta	ABS	NUMP	**abestu**	free	free
karguari	kargu	DAT	NUMS	**eutsi**	free	free
kargua	kargu	ABS	NUMS	**hartu**	col	col
kartzelaz	kartzela	INS	NUMS	**aldatu**	free	free
kartzelara	kartzela	ALA	NUMS	**bidali**	free	free
kartzelara	kartzela	ALA	NUMS	**eraman**	free	free
kasu	kasu	ABS	MG	**atzeman**	free	free
kasua	kasu	ABS	NUMS	**aztertu**	free	free
kasu	kasu	ABS	MG	**egin**	col	id
kasua	kasu	ABS	NUMS	**izan**	free	free
katalanez	katalan	INS	NUMP	**idatzi**	free	free
kea	ke	ABS	NUMS	**arnastu**	free	free
kereila	kereila	ABS	NUMS	**artxibatu**	free	free
kereila	kereila	ABS	NUMS	**jarri**	col	col
kezka	kezka	ABS	NUMS	**sortu**	col	col
kideak	kide	ABS	NUMP	**azaldu**	free	free
kide	kide	ABS	MG	**ukan**	free	free
kiratsa	kirats	ABS	NUMS	**jario**	col	col
kodea	kode	ABS	NUMS	**aldatu**	free	free
kokaina	kokaina	ABS	NUMS	**atzeman**	free	free
koktelak	koktel	ABS	NUMP	**jaurti**	free	free
kolpea	kolpe	ABS	NUMS	**hartu**	col	col
komunikabideak	komunikabide	ABS	NUMP	**itxi**	free	free
komunitatea	komunitate	ABS	NUMS	**osatu**	free	free
konfiantza	konfiantza	ABS	NUMS	**berreskura**	col	col
konfiantza	konfiantza	ABS	NUMS	**eduki**	col	col
konortea	konorte	ABS	NUMS	**galdu**	col	col
konponbidea	konponbide	ABS	NUMS	**eman**	col	col
konponbidean	konponbide	INE	NUMS	**jarri**	free	free
konpromisoa	konpromiso	ABS	NUMS	**berrets**	col	col
konpromisoak	konpromiso	ABS	NUMP	**bete**	col	col
konpromisoari	konpromiso	DAT	NUMS	**eutsi**	free	free
konpromisoa	konpromiso	ABS	NUMS	**izan**	free	free
konstituzioa	konstituzio	ABS	NUMS	**aldatu**	free	free
kontra	kontra	ABS	NUMS	**agertu**	col	col
kontraesana	kontraesan	ABS	NUMS	**iruditu**	free	free
kontratua	kontratu	ABS	NUMS	**amaitu**	free	free

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
kontrola	kontrol	ABS	NUMS	**bereganatu**	free	free
kontrolak	kontrol	ABS	NUMP	**egin**	free	free
kontrolak	kontrol	ABS	NUMP	**zorroztu**	free	free
kontseiluak	kontseilu	ABS	NUMP	**onartu**	free	free
kontura	kontu	ALA	NUMS	**aberastu**	free	free
kontzertua	kontzertu	ABS	NUMS	**egin**	col	col
kopara	kopa	ALA	NUMS	**sailkatu**	free	free
kopuruari	kopuru	DAT	NUMS	**egon**	free	free
kopurua	kopuru	ABS	NUMS	**emendatu**	free	free
kopurua	kopuru	ABS	NUMS	**hazi**	free	free
kopurua	kopuru	ABS	NUMS	**izan**	free	free
kopurua	kopuru	ABS	NUMS	**mugatu**	free	free
kopurua	kopuru	ABS	NUMS	**txikitu**	free	free
korrika	korrika	ABS	NUMS	**egin**	col	id
krimenez	krimen	INS	NUMP	**akusatu**	free	free
krimenak	krimen	ABS	NUMP	**egin**	free	free
krimenak	krimen	ABS	NUMP	**epaitu**	free	free
krimenak	krimen	ABS	NUMP	**ikertu**	free	free
krimentzat	krimen	PRO	MG	**jo**	free	free
kristalak	kristal	ABS	NUMP	**hautsi**	free	free
kutsadura	kutsadura	ABS	NUMS	**murriztu**	free	free
lagun	lagun	ABS	MG	**atxilotu**	free	free
lagunek	lagun	ERG	MG	**bisitatu**	free	free
lagunek	lagun	ERG	MG	**egin**	free	free
lagun	lagun	ABS	MG	**egon**	free	free
lagun	lagun	ABS	MG	**hil**	free	free
laguntzak	laguntza	ABS	NUMP	**banatu**	free	free
laguntza	laguntza	ABS	NUMS	**izan**	free	free
lagun	lagun	ABS	MG	**ukan**	col	col
lanak	lan	ABS	NUMP	**bildu**	free	free
lanean	lan	INE	NUMS	**egon**	col	col
lan	lan	ABS	MG	**eskatu**	free	free
langileek	langile	ERG	NUMP	**jasan**	free	free
lanetik	lan	ABL	NUMS	**irten**	free	free
lana	lan	ABS	NUMS	**izan**	free	free
lanean	lan	INE	NUMS	**jarraitu**	free	free
lankidetza	lankidetza	ABS	NUMS	**sustatu**	free	free
lana	lan	ABS	NUMS	**nabarmendu**	free	free
larritasunaz	larritasun	INS	NUMS	**ohartarazi**	free	free
larrua	larru	ABS	NUMS	**jo**	id	id
larunbatean	larunbat	INE	NUMS	**izan**	free	free
larunbatean	larunbat	INE	NUMS	**jokatu**	free	free
lasaitasuna	lasaitasun	ABS	NUMS	**eskatu**	free	free
lasterketa	lasterketa	ABS	NUMS	**irabazi**	free	free
legeak	lege	ABS	NUMP	**agindu**	free	free
legea	lege	ABS	NUMS	**bete**	col	col
legebiltzarra	legebiltzar	ABS	NUMS	**desegin**	col	col

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
legebiltzarrak	legebiltzar	ABS	NUMP	**onartu**	free	free
legea	lege	ABS	NUMS	**egin**	col	col
legea	lege	ABS	NUMS	**ezarri**	free	free
legea	lege	ABS	NUMS	**haustu**	col	col
legea	lege	ABS	NUMS	**izan**	free	free
legez	lege	INS	MG	**kanporatu**	col	col
lehendakariak	lehendakari	ERG	NUMS	**adierazi**	free	free
lehendakariarekin	lehendakari	SOZ	NUMS	**bildu**	free	free
lehentasuna	lehentasun	ABS	NUMS	**eman**	col	col
lehentasuna	lehentasun	ABS	NUMS	**ukan**	free	free
lehergailu	lehergailu	ABS	MG	**egin**	free	free
lehergailua	lehergailu	ABS	NUMS	**indargabetu**	free	free
lehian	lehia	INE	NUMS	**ibili**	free	free
lehia	lehia	ABS	NUMS	**izan**	free	free
lehian	lehia	INE	NUMS	**izan**	col	col
lehiaketara	lehiaketa	ALA	NUMS	**aurkeztu**	free	free
lehiaketa	lehiaketa	ABS	MG	**egin**	free	free
lehiaketa	lehiaketa	ABS	NUMS	**irabazi**	free	free
lehiaketa	lehiaketa	ABS	MG	**izan**	free	free
leihoa	leiho	ABS	NUMS	**ireki**	id	id
lekuz	leku	INS	MG	**aldatu**	col	col
lekua	leku	ABS	NUMS	**bete**	col	col
lekua	leku	ABS	NUMS	**egin**	col	col
leloa	lelo	ABS	NUMS	**aukeratu**	free	free
lelopean	lelope	INE	NUMS	**egin**	free	free
lerrook	lerro	ABS	PH	**idatzi**	free	free
liburua	liburu	ABS	NUMS	**argitaratu**	free	free
lidergoa	lidergo	ABS	NUMS	**eskuratu**	col	col
lider	lider	ABS	MG	**jarri**	col	col
liga	liga	ABS	NUMS	**irabazi**	free	free
lizentziak	lizentzia	ABS	NUMP	**gorde**	free	free
lotsa	lotsa	ABS	NUMS	**sentitu**	col	col
lotura	lotura	ABS	MG	**ukan**	col	col
lurrean	lur	INE	NUMS	**egon**	free	free
lurrean	lur	INE	NUMS	**eseri**	free	free
lurrean	lur	INE	NUMS	**etzan**	free	free
lurra	lur	ABS	NUMS	**jo**	id	id
lurpetik	lurpe	ABL	NUMS	**atera**	free	free
lurra	lur	ABS	NUMS	**ukitu**	id	id
mahaian	mahai	INE	NUMS	**eseri**	col	col
maila	maila	ABS	NUMS	**galdu**	col	col
maillota	maillot	ABS	NUMS	**jantzi**	free	free
maisutasuna	maisutasun	ABS	NUMS	**erakutsi**	free	free
malkoak	malko	ABS	NUMP	**isuri**	col	col
mamuak	mamu	ABS	NUMP	**uxatu**	id	id
manifestazioa	manifestazio	ABS	NUMS	**antolatu**	free	free
manifestaziora	manifestazio	ALA	NUMS	**deitu**	free	free

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
manifestua	manifestu	ABS	NUMS	**sinatu**	free	free
marka	marka	ABS	NUMS	**ezarri**	col	col
markagailuan	markagailu	INE	NUMS	**aurreratu**	col	col
markagailuari	markagailu	DAT	NUMS	**eutsi**	free	free
markagailuan	markagailu	INE	NUMS	**hurbildu**	free	free
markoa	marko	ABS	NUMS	**gainditu**	col	col
martxan	martxa	INE	NUMS	**egon**	col	col
martxan	martxa	INE	NUMS	**ipini**	col	col
martxan	martxa	INE	NUMS	**izan**	col	col
meategia	meategi	ABS	NUMS	**ustiatu**	free	free
mehatxuak	mehatxu	ABS	NUMP	**jaso**	col	col
mendean	mende	INE	NUMS	**hartu**	col	col
menditik	mendi	ABL	NUMS	**jaitsi**	free	free
mendira	mendi	ALA	NUMS	**joan**	free	free
merkantziak	merkantzia	ABS	NUMP	**garraiatu**	free	free
merkatua	merkatu	ABS	NUMS	**zabaldu**	free	free
mesede	mesede	ABS	MG	**egin**	col	id
metrotik	metro	ABL	NUMS	**erori**	free	free
mezua	mezu	ABS	NUMS	**bidali**	free	free
mezua	mezu	ABS	NUMS	**izan**	free	free
miaketak	miaketa	ABS	NUMP	**egin**	col	col
milioi	milioi	ABS	MG	**inbertitu**	free	free
mina	mina	ABS	NUMS	**sentitu**	free	free
ministerioak	ministerio	ABS	NUMP	**jakinarazi**	free	free
ministroak	ministro	ERG	NUMS	**adierazi**	free	free
ministroarekin	ministro	SOZ	NUMS	**bildu**	free	free
ministroak	ministro	ABS	NUMP	**esan**	free	free
mobilizazioei	mobilizazio	DAT	NUMP	**ekin**	free	free
mobilizazioekin	mobilizazio	SOZ	NUMP	**jarraitu**	free	free
modua	modu	ABS	NUMS	**aldatu**	free	free
moduan	modu	INE	NUMS	**aritu**	free	free
moduan	modu	INE	NUMS	**egon**	free	free
moduan	modu	INE	NUMS	**funtzionatu**	free	free
modua	modu	ABS	NUMS	**izan**	free	free
modura	modu	ALA	NUMS	**ulertu**	free	free
motorrak	motor	ABS	NUMP	**berotu**	id	id
mugak	muga	ABS	NUMP	**ezabatu**	col	col
mugak	muga	ABS	NUMP	**gainditu**	col	col
muga	muga	ABS	NUMS	**izan**	col	col
mugimendua	mugimendu	ABS	NUMS	**kriminaliza**	free	free
mundua	mundu	ABS	NUMS	**aldatu**	free	free
munduan	mundu	INE	NUMS	**barneratu**	free	free
mundua	mundu	ABS	NUMS	**euskaldundu**	free	free
mundua	mundu	ABS	NUMS	**salbatu**	free	free
museoa	museo	ABS	NUMS	**kudeatu**	free	free
musika	musika	ABS	NUMS	**egin**	col	col
musika	musika	ABS	NUMS	**izan**	free	free

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
muturreraino	mutur	ABU	NUMS	**eraman**	id	id
mutxikoak	mutxiko	ABS	NUMP	**dantzatu**	free	free
muzin	muzin	ABS	MG	**egin**	col	id
nagusitasuna	nagusitasun	ABS	NUMS	**erakutsi**	col	col
nahasmena	nahasmen	ABS	NUMS	**sortu**	free	free
nahiari	nahi	DAT	NUMS	**erantzun**	free	free
nahian	nahi	INE	NUMS	**ibili**	col	col
nahietara	nahi	ALA	NUMP	**makurtu**	col	col
nahi	nahi	ABS	MG	**ukan**	col	id
natura	natura	ABS	NUMS	**babestu**	free	free
nazioartean	nazioarte	INE	NUMS	**hedatu**	free	free
nazioartean	nazioarte	INE	NUMS	**lehiatu**	free	free
naziotasuna	naziotasun	ABS	NUMS	**aldarrikatu**	free	free
negoziazioak	negoziazio	ABS	NUMP	**blokeatu**	free	free
nekea	neke	ABS	NUMS	**nabaritu**	free	free
obran	obra	INE	NUMS	**oinarritu**	free	free
odola	odol	ABS	NUMS	**isuri**	col	col
ofizialtasuna	ofizialtasun	ABS	NUMS	**aitortu**	free	free
ofizialtasuna	ofizialtasun	ABS	NUMS	**aldarrikatu**	free	free
oharrean	ohar	INE	NUMS	**bidali**	free	free
ohitura	ohitura	ABS	NUMS	**bihurtu**	free	free
ohitura	ohitura	ABS	NUMS	**bilakatu**	col	col
oholtzara	oholtza	ALA	NUMS	**igo**	col	col
oinarriak	oinarri	ABS	NUMP	**finkatu**	free	free
oinarritzat	oinarri	PRO	MG	**hartu**	col	col
omenez	omen	INS	NUMP	**egin**	free	free
onetik	on	ABL	NUMS	**atera**	id	id
ondare	ondare	ABS	MG	**izendatu**	free	free
ondoan	ondo	INE	NUMS	**egon**	free	free
ondoan	ondo	INE	NUMS	**izan**	free	free
ondorioak	ondorio	ABS	NUMP	**atera**	col	col
ondorioak	ondorio	ABS	NUMP	**pairatu**	col	col
ondorioz	ondorio	INS	MG	**zendu**	free	free
ondorioz	ondorio	INS	MG	**zigortu**	free	free
onespena	onespen	ABS	NUMS	**jaso**	col	col
opa	opa	ABS	NUMS	**ukan**	col	id
operazioan	operazio	INE	NUMS	**atxilotu**	free	free
operazioarekin	operazio	SOZ	NUMS	**lotu**	free	free
oporretara	opor	ALA	NUMP	**joan**	free	free
ordez	orde	INS	MG	**ari_izan**	free	free
ordezkariak	ordezkari	ABS	NUMP	**bildu**	free	free
ordu	ordu	ABS	MG	**iraun**	free	free
orkatila	orkatila	ABS	NUMS	**bihurritu**	col	col
ospitaletik	ospitale	ABL	NUMS	**atera**	free	free
ospitalean	ospitale	INE	NUMS	**egon**	free	free
ostean	oste	INE	NUMS	**etorri**	free	free
ostiralean	ostiral	INE	NUMS	**egin**	free	free

Bigrama normalizatua					**Sailkapena**	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
palestinarrekin	palestinar	SOZ	NUMP	**negoziatu**	free	free
panderoa	pandero	ABS	NUMS	**jo**	free	free
pankartari	pankarta	DAT	NUMS	**eutsi**	free	free
paretik	pare	ABL	NUMS	**igaro**	free	free
parekoa	pare	GELABS	NUMS	**izan**	free	free
parkea	parke	ABS	NUMS	**eraiki**	free	free
parlamentua	parlamentu	ABS	NUMS	**desegin**	col	col
parlamentuan	parlamentu	INE	NUMS	**eztabaidatu**	free	free
partaide	partaide	ABS	MG	**izan**	free	free
parte-hartzea	parte-hartze	ABS	NUMS	**bultzatu**	free	free
partidak	partida	ERG	MG	**galdu**	free	free
partida	partida	ABS	MG	**irabazi**	free	free
partidak	partida	ERG	MG	**irabazi**	free	free
partida	partida	ABS	MG	**izan**	free	free
partida	partida	ABS	MG	**jokatu**	free	free
partidua	partidu	ABS	NUMS	**jokatu**	free	free
paseoan	paseo	INE	NUMS	**ibili**	col	col
pausoa	pauso	ABS	NUMS	**izan**	free	free
pedalei	pedal	DAT	NUMP	**eragin**	col	col
pertsona	pertsona	ABS	MG	**egon**	free	free
pertsonak	pertsona	ABS	NUMP	**izan**	free	free
pertsonei	pertsona	DAT	NUMP	**lagundu**	free	free
pertsona	pertsona	ABS	MG	**zauritu**	free	free
petrolioa	petrolio	ABS	NUMS	**garestitu**	free	free
pezeta	pezeta	ABS	MG	**balio_izan**	free	free
pezeta	pezeta	ABS	MG	**bideratu**	free	free
pezeta	pezeta	ABS	MG	**jaso**	free	free
pianoa	piano	ABS	NUMS	**jo**	free	free
piezak	pieza	ABS	NUMP	**interpretatu**	free	free
pilotalekuan	pilotaleku	INE	NUMS	**jokatu**	free	free
pilotaria	pilotari	ABS	NUMS	**izan**	free	free
pintxoak	pintxo	ABS	NUMP	**jan**	free	free
pistan	pista	INE	NUMS	**ari_izan**	free	free
plana	plan	ABS	NUMS	**egin**	col	col
planta	planta	ABS	NUMS	**egin**	col	col
plazan	plaza	INE	NUMS	**egon**	free	free
plazan	plaza	INE	NUMS	**izan**	free	free
poemak	poema	ABS	NUMP	**idatzi**	free	free
politikaria	politikari	ABS	NUMS	**jaio**	free	free
porrota	porrot	ABS	NUMS	**izan**	col	col
postuan	postu	INE	NUMS	**bukatu**	free	free
postuan	postu	INE	NUMS	**egon**	free	free
postua	postu	ABS	NUMS	**eskuratu**	free	free
postuari	postu	DAT	NUMS	**eutsi**	col	col
postuan	postu	INE	NUMS	**izan**	free	free
postuan	postu	INE	NUMS	**jarraitu**	free	free
postura	postura	ABS	MG	**igo**	free	free

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
postura	postura	ABS	MG	**jaitsi**	free	free
poza	poz	ABS	NUMS	**agertu**	free	free
premiei	premia	DAT	NUMP	**erantzun**	col	col
prentsaurrekoan	prentsaurreko	INE	NUMS	**eman**	free	free
presentzia	presentzia	ABS	NUMS	**areagotu**	free	free
presidenteak	presidente	ABS	NUMP	**esan**	free	free
presidente	presidente	ABS	MG	**izan**	free	free
presidente	presidente	ABS	MG	**izendatu**	free	free
presidenteak	presidente	ERG	NUMS	**jakinarazi**	free	free
preso	preso	ABS	MG	**hartu**	col	col
presoak	preso	ABS	NUMP	**torturatu**	free	free
prezioan	prezio	INE	NUMS	**saldu**	free	free
proba	proba	ABS	NUMS	**irabazi**	free	free
proba	proba	ABS	MG	**izan**	free	free
proba	proba	ABS	MG	**jokatu**	free	free
produktuak	produktu	ABS	NUMP	**dastatu**	free	free
profesionalak	profesional	ABS	NUMP	**izan**	free	free
profesionaletara	profesional	ALA	NUMP	**pasatu**	free	free
proiektua	proiektu	ABS	NUMS	**aurkeztu**	free	free
proiektua	proiektu	ABS	NUMS	**inposatu**	free	free
proiektuan	proiektu	INE	NUMS	**sinetsi**	free	free
promesa	promes	ABS	NUMS	**egin**	col	col
proposamena	proposamen	ABS	NUMS	**aurkeztu**	free	free
proposamena	proposamen	ABS	NUMS	**izan**	free	free
protestara	protesta	ALA	NUMS	**deitu**	free	free
protokoloa	protokolo	ABS	NUMS	**ordezkatu**	free	free
prozesua	prozesu	ABS	NUMS	**amaitu**	free	free
prozesua	prozesu	ABS	NUMS	**bultzatu**	col	col
prozesua	prozesu	ABS	NUMS	**oztopatu**	free	free
puntura	puntu	ALA	NUMS	**gerturatu**	free	free
puntuak	puntu	ABS	NUMP	**pilatu**	free	free
sagardoa	sagardo	ABS	NUMS	**dastatu**	free	free
sailetan	sail	INE	NUMP	**banatu**	free	free
sailak	sail	ERG	NUMS	**jakinarazi**	free	free
sailkapena	sailkapen	ABS	NUMS	**lortu**	free	free
saioa	saio	ABS	NUMS	**aurkeztu**	free	free
saioa	saio	ABS	NUMS	**eskaini**	col	col
saioa	saio	ABS	NUMS	**estreinatu**	free	free
saioa	saio	ABS	NUMS	**jokatu**	free	free
salaketa	salaketa	ABS	NUMS	**jarri**	col	col
salbuespena	salbuespen	ABS	NUMS	**izan**	free	free
sareetara	sare	ALA	NUMP	**bidali**	id	id
sarietarako	sari	ALAGEL	NUMP	**izendatu**	free	free
sasoia	sasoi	ABS	NUMS	**amaitu**	free	free
segundo	segundo	ABS	MG	**atera**	col	col
segundoan	segundo	INE	NUMS	**ebaki**	free	free
segundo	segundo	ABS	MG	**jan**	col	col

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
segundo	segundo	ABS	MG	**kendu**	free	free
sekretupetik	sekretupe	ABL	NUMS	**atera**	col	col
seme-alabak	seme-alabak	ABS	NUMP	**zaindu**	free	free
semea	seme	ABS	NUMS	**izan**	free	free
senideekin	senide	SOZ	NUMP	**bildu**	free	free
sistema	sistema	ABS	NUMS	**ezarri**	free	free
sistemara	sistema	ALA	NUMS	**itzuli**	free	free
soldadu	soldadu	ABS	MG	**hil**	free	free
sormena	sormen	ABS	NUMS	**landu**	free	free
sostengua	sostengu	ABS	NUMS	**adierazi**	col	col
sostengua	sostengu	ABS	NUMS	**eman**	col	col
su-etena	su-eten	ABS	NUMS	**aldarrikatu**	free	free
suetena	sueten	ABS	NUMS	**hautsi**	col	col
su-etena	su-eten	ABS	NUMS	**hitzartu**	free	free
suhiltzaileek	suhiltzaile	ERG	NUMP	**itzali**	free	free
sua	su	ABS	NUMS	**piztu**	id	id
susmoa	susmo	ABS	NUMS	**ukan**	col	col
taldearekin	talde	SOZ	NUMS	**alderatu**	free	free
taldetik	talde	ABL	NUMS	**bota**	free	free
taldeak	talde	ABS	NUMP	**egin**	free	free
taldera	talde	ALA	NUMS	**egokitu**	free	free
talde	talde	ABS	MG	**egon**	free	free
taldetik	talde	ABL	NUMS	**etorri**	free	free
taldea	talde	ABS	NUMS	**indartu**	free	free
taldearekin	talde	SOZ	NUMS	**korritu**	free	free
taldea	talde	ABS	NUMS	**sortu**	free	free
tanto	tanto	ABS	MG	**egin**	col	col
tartea	tarte	ABS	NUMS	**ireki**	col	col
tartea	tarte	ABS	MG	**utzi**	free	free
telebista	telebista	ABS	NUMS	**piztu**	free	free
telefonora	telefono	ALA	NUMS	**deitu**	free	free
tenorea	tenore	ABS	NUMS	**heldu**	free	free
tentazioan	tentazio	INE	NUMS	**erori**	col	col
tentsioa	tentsio	ABS	NUMS	**mantendu**	free	free
terrorismoa	terrorismo	ABS	NUMS	**garaitu**	free	free
titular	titular	ABS	MG	**izan**	free	free
titulua	titulu	ABS	NUMS	**lortu**	free	free
tontorra	tontor	ABS	NUMS	**zapaldu**	col	col
tradizioari	tradizio	DAT	NUMS	**eutsi**	free	free
trafikoan	trafiko	INE	NUMS	**jardun**	free	free
traineru	traineru	ABS	MG	**lehiatu**	free	free
tramitera	tramite	ALA	NUMS	**onartu**	col	col
trenera	tren	ALA	NUMS	**igo**	id	id
tunela	tunel	ABS	NUMS	**zulatu**	free	free
turismoa	turismo	ABS	NUMS	**sustatu**	free	free
turistak	turista	ABS	NUMP	**erakarri**	free	free
txakolina	txakolin	ABS	NUMS	**dastatu**	free	free

Bigrama normalizatua					Sailkapena	
izena				aditza	(1)	(2)
forma	lema	kasua	mugat.			
txandatan	txanda	INE	MG	**banatu**	free	free
txanda	txanda	ABS	NUMS	**izan**	free	free
txapelduna	txapeldun	ABS	NUMS	**izan**	free	free
txapeldunorde	txapeldunorde	ABS	MG	**izan**	free	free
txapela	txapel	ABS	NUMS	**jantzi**	id	id
txapelketa	txapelketa	ABS	NUMS	**antolatu**	free	free
txapelketa	txapelketa	ABS	NUMS	**izan**	free	free
txapelketarako	txapelketa	ALAGEL	NUMS	**sailkatu**	free	free
txapela	txapel	ABS	NUMS	**lortu**	id	id
txinga	txinga	ABS	NUMS	**eroan**	col	col
txirrindularia	txirrindulari	ABS	NUMS	**izan**	free	free
txupinazoa	txupinazo	ABS	NUMS	**bota**	free	free
udalak	udal	ABS	NUMP	**desegin**	free	free
udaltzaingoak	udaltzaingo	ABS	NUMP	**jakinarazi**	free	free
ukituak	ukitu	ABS	NUMP	**eman**	col	col
umezurtz	umezurtz	ABS	MG	**geratu**	col	col
unibertsitatean	unibertsitate	INE	NUMS	**ikasi**	free	free
uranioa	uranio	ABS	NUMS	**aberastu**	col	col
ura	ur	ABS	NUMS	**berotu**	free	free
ura	ur	ABS	NUMS	**izan**	free	free
urratsak	urrats	ABS	NUMP	**egin**	col	col
urratsak	urrats	ABS	NUMP	**eman**	col	col
urtez	urte	INS	MG	**aritu**	free	free
urtebetez	urtebete	INS	MG	**luzatu**	free	free
urteetan	urte	INE	NUMP	**gertatu**	free	free
urtean	urte	INE	NUMS	**hasi**	free	free
urte	urte	ABS	MG	**igaro**	free	free
urte	urte	ABS	MG	**iraun**	free	free
urte	urte	ABS	MG	**izan**	free	free
urte	urte	ABS	MG	**ukan**	free	free
urteurrena	urteurren	ABS	NUMS	**gogoratu**	free	free
ustekabea	ustekabe	ABS	NUMS	**izan**	free	free
xedea	xede	ABS	NUMS	**ukan**	free	free
xehetasunak	xehetasun	ABS	NUMP	**eman**	col	col
zaborrak	zabor	ABS	NUMP	**kudeatu**	free	free
zalantza	zalantza	ABS	NUMS	**izan**	col	col
zalantzak	zalantza	ABS	NUMP	**uxatu**	col	col
zatiketa	zatiketa	ABS	NUMS	**gainditu**	free	free
zerbitzura	zerbitzu	ALA	NUMS	**egon**	col	col
zerbitzua	zerbitzu	ABS	NUMS	**eman**	col	col
zerbitzuak	zerbitzu	ABS	NUMP	**pribatizatu**	free	free
zergak	zerga	ABS	NUMP	**igo**	free	free
zera	zer	ABS	NUMS	**galdetu**	free	free
zergak	zerga	ABS	NUMP	**ordaindu**	free	free
zera	zer	ABS	NUMS	**izan**	free	free
zerrendak	zerrenda	ABS	NUMP	**aurkeztu**	free	free
zerura	zeru	ALA	NUMS	**begiratu**	free	free

Bigrama normalizatua					Sailkapena	
izena				**aditza**	(1)	(2)
forma	lema	kasua	mugat.			
zerua	zeru	ABS	NUMS	**ukitu**	id	id
ziegatik	ziega	ABL	NUMS	**atera**	free	free
zigarroa	zigarro	ABS	NUMS	**erre**	free	free
zigorgabetasuna	zigorgabetasun	ABS	NUMS	**amaitu**	free	free
zigorra	zigor	ABS	NUMS	**jaso**	col	col
zinemagilea	zinemagile	ABS	NUMS	**jaio**	free	free
zorra	zor	ABS	NUMS	**ezabatu**	free	free
zor	zor	ABS	MG	**izan**	col	id
zorra	zor	ABS	NUMS	**kitatu**	col	col
zuloa	zulo	ABS	NUMS	**egin**	id	id
zutabea	zutabe	ABS	NUMS	**idatzi**	free	free
zutoinera	zutoin	ALA	NUMS	**bidali**	col	col
zutoina	zutoin	ABS	NUMS	**jo**	free	free
zuzenbidea	zuzenbide	ABS	NUMS	**ikasi**	free	free
zuzendariarekin	zuzendari	SOZ	NUMS	**bildu**	free	free
zuzendariak	zuzendari	ABS	NUMP	**esan**	free	free
zuzenketak	zuzenketa	ABS	NUMP	**aurkeztu**	free	free

B. ERANSKINA

Karakterizazio-emaitzaren erakusgarria

Idiomatikotasunaren ranking idealarekiko Kendall τ_B koefizientearen araberako hein-korrelazio onena duen neurriak (DSim_L2_Indri_hit_erl_NV) sortutako rankingaren hiru erakusgarri antolatu ditugu, hurrenez hurren, hasierako, erdialdeko eta amaierako tarteetakoak.

DSim_L2_Indri_hit_erl_NV neurriak konbinazioan konposizionaltasun semantikoa neurtzen du, antzekotasun distribuzionalaren bidez. Neurri honen xehetasunak VII.1.2 atalean eman ditugu; zehazki, VII.1.2.4 azpiatalean.

Taula bakoitzean 35 hautagai daude.

- B.1 Rankingeko 1-35 UF hautagaiak. Rankingaren lehen posizioetan, `id` kategoriako UF hautagaiak ugari agertzen dira, continuumarekiko korrelazio handiaren zantzua dena.
- B.2 Rankingeko 200-234 UF hautagaiak. Rankingaren posizio horietan, UF mota ugariena `col` kategoria izatea da espero behar dena.
- B.3 Rankingeko 700-734 UF hautagaiak. `free` kategoriako hautagaiak dira ugarienak.

Itxaropen horien arabera ranking-tarte bakoitzean ugarien behar luketen hautagaien errenkadak hondo grisez nabarmendu ditugu tauletan.

B.1 Rankingeko 1-35 UF hautagaiak

ord.	bigrama	mota	DSim_L2_Indri_hit_erl_NV
1	legez kanporatu	col	0,07073
2	ametsa bete	col	0,18169
3	lerrook idatzi	free	0,18354
4	erditik kendu	id	0,18889
5	bazterrak harrotu	id	0,19697
6	adarra jo	id	0,20766
7	bazterrak nahastu	id	0,21429
8	hondoa jo	id	0,22025
9	hautsak harrotu	id	0,22615
10	bazterrean utzi	id	0,25089
11	egurra eman	id	0,25621
12	garrantzia kendu	col	0,25648
13	aurrera eraman	id	0,27450
14	eskutik etorri	id	0,27809
15	bidea eman	id	0,28132
16	aurpegia garbitu	free	0,28704
17	jardunaldiak egin	free	0,28905
18	eskutik heldu	id	0,29730
19	arriskuak hartu	col	0,29745
20	atzera egin	col	0,29917
21	bidetik joan	col	0,29963
22	hanka sartu	id	0,30667
23	gorriak ikusi	id	0,30814
24	gogoan hartu	col	0,30948
25	poza agertu	free	0,30952
26	elkarlanean aritu	free	0,31069
27	burua zuritu	id	0,31395
28	lagun ukan	col	0,31899
29	indarra galdu	free	0,31951
30	errotik moztu	id	0,32099
31	odola isuri	col	0,32105
32	izenarekin ezagutu	free	0,32192
33	bizia galdu	col	0,32262
34	salbuespena izan	free	0,32345
35	arriskuan ikusi	col	0,32488

B.2 Rankingeko 200-234 UF hautagaiak

ord.	bigrama	mota	DSim_L2_Indri_hit_erl_NV
200	promesa egin	col	0,42327
201	kanpotik ekarri	free	0,42330
202	arazorik eduki	free	0,42368
203	gainean egon	id	0,42433
204	alboan izan	id	0,42481
205	zalantza izan	col	0,42520
206	lehentasuna eman	col	0,42538
207	susmoa ukan	col	0,42570
208	lan eskatu	free	0,42595
209	jarraipena izan	free	0,42642
210	zuzendariak esan	free	0,42652
211	nahasmena sortu	free	0,42701
212	erantzuna jaso	col	0,42760
213	eztabaida izan	col	0,42785
214	emaztea izan	free	0,42785
215	zerbitzura egon	col	0,42881
216	aitzindaria izan	free	0,42961
217	porrota izan	col	0,42996
218	atzetik joan	id	0,42998
219	muga izan	col	0,43002
220	urratsak egin	col	0,43025
221	ezinbestekotzat jo	col	0,43119
222	aldaketak egin	col	0,43253
223	mugak gainditu	col	0,43258
224	xehetasunak eman	col	0,43261
225	indarrak neurtu	id	0,43271
226	parekoa izan	free	0,43301
227	lurra jo	id	0,43349
228	arnasa hartu	col	0,43410
229	moduan egon	free	0,43425
230	ironia erabili	free	0,43478
231	eskaera izan	free	0,43498
232	urteetan gertatu	free	0,43575
233	taldeak egin	free	0,43628
234	giltza izan	id	0,43662

B.3 Rankingeko 700-734 UF hautagaiak

ord.	bigrama	mota	DSim_L2_Indri_hit_erl_NV
700	bainua hartu	col	0,63190
701	euro eman	free	0,63307
702	informazioa trukatu	free	0,63333
703	sostengua adierazi	col	0,63370
704	kartzelaz aldatu	free	0,63415
705	jendea hunkitu	free	0,63415
706	kopurua hazi	free	0,63462
707	adostasuna lortu	col	0,63473
708	helbidera idatzi	free	0,63636
709	harresia eraiki	id	0,63740
710	hasieran aurreikusi	free	0,63855
711	ministroak esan	free	0,63961
712	sistemara itzuli	free	0,64063
713	irudimena landu	free	0,64103
714	konstituzioa aldatu	free	0,64138
715	interesak defenditu	free	0,64173
716	konfiantza berreskura	col	0,64179
717	buru-belarri murgildu	free	0,64286
718	sailetan banatu	free	0,64286
719	etxean atxiki	free	0,64286
720	etorkizuna eraiki	id	0,64423
721	taldearekin alderatu	free	0,64530
722	festan murgildu	col	0,64545
723	arte luzatu	free	0,64569
724	indarkeriarekin lotu	free	0,64674
725	gaindi ibili	free	0,64706
726	esperientziak partekatu	free	0,64754
727	etorkizunaz erabaki	free	0,64881
728	finalerdietara iritsi	free	0,64908
729	arte atzeratu	free	0,65027
730	errealitatea islatu	free	0,65038
731	harremanak haustu	col	0,65116
732	lehendakariak adierazi	free	0,65116
733	zorra ezabatu	free	0,65179
734	elkartasuna adierazi	col	0,65217

C. ERANSKINA

Elhuyar Web-corpusen atariko "Hitz-konbinazioak" atalaren erakusgarria

Tesi-lan honetan `izena+aditza` osaerako konbinazioak forma kanonikoan erauzteko garatutako teknikak egokitu ditugu osaera desberdineko bi konbinazio-mota erauzteko: `izena+izenondoa` eta `izena+izena` konbinazioak. Horiek erabiliz, Elhuyar Fundazioaren *Web-corpusen Atari*ko euskarazko corpus elebakarra prozesatu da (125 miloi hitz), eta erauzitako hautagaien agerkidetzaren bidezko karakterizazioa egin da, zenbait AM elkartze-neurri kalkulatuz. Emaitzak doan kontsulta daitezke atariko "Hitz-konbinazioak" atalean (http://webcorpusak.elhuyar.org/cgi-bin/kolokatuak.py).

Egindako lanaren erakusgarri, aipatu hiru osaeretako hamarna adibide erakutsi ditugu hurrengo hiru tauletan, t neurriaren araberako rankingeko lehen hamarrak hain zuzen ere. f bigramaren maiztasuna da, f_1 izenaren bigramen maiztasuna, eta f_2, aditzaren bigramena.

C.1 izena+aditza konbinazioak

bigrama	f	f_1	f_2	t	adibidea
adostasuna lortu	646	1592	44605	24,94	*Horren ondoren, Jaurlaritzaren, aldundien eta udalen artean **lortutako adostasuna** azpimarratu zuen.*
adostasunera iritsi	185	252	16670	13,55	*Hori mahai gainean jartzen denean, oker egin dela eta gehiengo demokratikoak plenoan planteatzen dituen gauzak errespetatuko direnean, ez da inongo arazorik izango **adostasunera iristeko**.*
adostasuna bilatu	185	1592	9352	13,42	*Parte-hartzezko aurrekontuak tokiko gobernuan herritarren parte-hartzea sustatzeko mekanismotzat sortu ziren, bereziki, gastu publikoarekin lotutako gaietan; hau da, mekanismo horrek agintarien eta herritarren arteko **adostasuna bilatzen** du, udaletxeko gastuak definitzeko eta aurrekontuen gauzapenaren jarraipena egiteko.*
adostasuna azaldu	153	1592	21863	11,89	*Euskararen garapenerako zerbitzuak proposamena eta letratuak txostena egin dute, bai eta Kontuhartzailetza delegatuak bere **adostasuna azaldu** ere.*
adostasuna adierazi	91	1592	19877	8,98	*Horrenbestez, txostenerako helarazitako agindu proiektuaren testuarekiko bere **adostasuna adierazita** amaitzen du.*
adostasunera heldu	67	252	9370	8,14	*Ez da bidezkoa alderdiak isilbidezko **adostasunera heltzea**.*
adostasunean oinarritu	56	68	10722	7,47	***Adostasunean oinarriturik**, sakoneko herri-eredu aldaketa bultzatu du ezker abertzaleak, EAk eta Aralarrek osatzen zuten udal gobernuak.*
adostasuna egon	145	1592	229583	6,89	*Eta oro har, hezkuntzan adibidez, **adostasuna egon** delako lortu dira lortutakoak.*
adostasuna agertu	52	1592	25019	6,27	*Hirugarrena: hondartzainaren kontratazinorako hautaketa prozesua arautuko daben oinarriak langileen ordezkariakaz eztabaidatu dira, eta euren **adostasuna agertu** dabe.*
adostasuna eskatu	51	1592	30424	5,99	*Zuzenbidea eskatu ahal da, baina **adostasuna eskatu** ahal izateko besteek berau inspiratzen duten printzipioei men egin behar diete.*

C.2 izena+izenondoa konbinazioak

bigrama	f	f_1	f_2	t	adibidea
adostasun zabal	228	658	15223	14,78	*Egia da, praktikan, hizkuntzaren molde eta erabilera oinarrizkoenen egokitasunari eta gramatikaltasunari buruzko **adostasun zabal** samarra lortua dugula, baina, hizkuntzaren kodebalio oinarrizkoena bermatzen duen adostasungune finko horretatik urrundu ahala, lausotuz eta zalantzagarriago bilakatuz joaten dira gauzak.*
adostasun handi	172	658	142601	9,64	*Inkestan galde egin delarik egiazko euskal herritar izateko euskara jakin behar denetz iritzi ezberdinak agertu dira, **adostasun handiena** iparraldean delarik eta ezadostasun handiena autonomi elkartean.*
adostasun sozial	64	658	26952	6,92	*Neskek, biharko emakumeek, zaintzalanean konpartituriko errespontsabilitatea lortuko dute ala ez, haien aukera pertsonalaren ondorioz berriz ere, **adostasun sozialik** gabe.*
adostasun politiko	66	658	30905	6,91	*Aipamen berezia egin behar zaio prestazioen eta zerbitzuen zorroari buruzko dekretuari; izan ere, garapen teknikoarekin eta **adostasun politikoarekin** loturiko lan handiagoa eskatzen du, maila instituzional bakoitzak onartu behar baitu dagokion alderdian.*
adostasun etiko	34	658	1581	5,74	*Bigarren eztabaidaren helburua lehenbizikoaren berdina izan da: **adostasun etiko** batera biltzea nortasun erlijioso ezberdinetakoak direnak ez ezik inongo fede edo sinesmen erlijiosorik ez dutenak ere.*
adostasun minimo	11	658	409	3,28	*Irakasleen artean badago **adostasun minimo** hori lortzeko kezka.*
adostasun oso	37	658	65689	2,63	*Bestalde, **adostasun osoa** dago gizarte mugimenduen artean beste mundu bat eraikitzeko inperialismoa behinik eta betirako baztertu behar dela.*
adostasun orokor	15	658	44096	0,23	*Printzipio horiek onartzen ez dituzten taldeak edo ideologiak bakarrik utzi ditzake gizarteak **adostasun orokorretik** kanpo.*
adostasun berri	17	658	107459	-4,21	*Ez da konpromiso sendorik hartu, baina bai beroketaren aurkako **adostasun berri** batera iristeko oinarriak, 2013tik aurrera Kiotokoa ordeztuko duen ituna diseinatzeko bide direnak.*

C.3 izena+izena konbinazioak

bigrama	f	f_1	f_2	t	adibidea
adostasun-maila	126	344	49654	10,60	*1982an Euskararen Legearen inguruan lortutako **adostasun maila** bederen lortu beharko genuke.*
adostasun ez	46	344	7191	6,63	*Labelaz gain kalitatea zertifikatzen duten beste bost marka martxan egoteak erakusten du ekoizleen artean gaur egun dagoen **adostasun eza**.*
adostasun-puntu	36	344	3740	5,91	*Europako urak betiko lege baten pean egotea nahi baldin badugu, europar komunitateetako justizia auzitegiaren paper nagusiarekin jolastu beharko da, Europako uraren direktibarekin batera eta estatu kideen laguntzarekin **adostasun-puntu** erdi batera iristeko.*
adostasun-prozesu	23	344	16449	4,31	*Zenbait filosoforen ustez, horixe besterik ez da etika, hots, **adostasun-prozesu** bat.*
adostasun-txosten	15	344	2487	3,78	*Nafarroako Ingurugiro Kontseiluak 1996.eko ekainaren 20an Planaren Proiektuari buruzko adostasun txostena eman zuen.*
adostasun falta	15	344	3726	3,74	*Agian beste oinarrizko **adostasun faltaren** adierazgarri besterik ez da, baina honek hizkuntza politikari berari ez dagokion zama ezartzea ere esan nahi du, inolako mesederik egiten ez dion zama.*
adostasun-batzar	14	344	5395	3,54	*Lea Artibai ikastetxea 5 S metodologiaren ezarpena 2 **adostasun batzarra** 1 fasea argitu eta burutze jarraibideak ezarri jarraian agertzen diren kontzeptuak argitzen saiatzen gara.*
adostasun-proba	12	344	2043	3,38	*Aipagarria da, zentzu horretan, hondakinen kontrolerako bi fase berri txertatu direla: alde batetik, hondakinen oinarrizko ezaugarrien berri jasotzeko probak eta, bestetik, **adostasun probak**.*
adostasun-agiri	12	344	4263	3,29	*Bi alderdien erantzuna ikusirik, hasierako saio bat egingo da, alderdi bakoitzarekin, bitartekotzan parte hartzeko ados daudela adieraz dezaten; horretarako, **adostasun-agiri** bat sinatuko dute.*
adostasun-adierazpen	11	344	3188	3,18	*Zatia: **adostasun adierazpena**: europar hiriak garapen jasangarrirantz.*