Idiomatikotasunaren karakterizazio automatikoa: *izena + aditza* konbinazioak

*Antton Gurrutxaga**

Hizkuntza eta Teknologia saila, Elhuyar Fundazioa

Iñaki Alegria, Xabier Artola

Ixa taldea-Informatika Fakultatea, UPV/EHU.

* a.gurrutxaga@elhuyar.com

DOI: 10.1387/ekaia.14544

Jasoa: 2015-05-26 Onartua: 2015-12-22

Laburpena: Ikerkuntza-lan honen helburua izan da izena+aditza osaerako unitate fraseologikoak (UFak) corpusetik automatikoki eskuratzeko eta haien idiomatikotasunaren arabera karakterizatzeko teknikak garatzea eta esperimentalki testatzea. Idiomatikotasuna UFen ezaugarri definitzailetzat hartu dugu, eta haren lau propietate hauek hartu ditugu neurgai gisa: instituzionalizazioa (idiosinkrasia estatistikoa), ez-konposizionaltasun semantikoa, finkapen morfosintaktikoa eta finkapen lexikala. Ondorio nagusia da arlo honetan estandar diren agerkidetza-tekniken emaitzak modu esanguratsuan gainditu direla, batez ere teknika semantikoen bidez, baina baita malgutasun morfosintaktikoaren neurketen bidez ere. Aldiz, malgutasun lexikalaren neurketek ez dute espero izatekoa zen mailako emaitza izan. Azkenik, teoria fraseologikoaren aurresan batzuen ebidentzia esperimentalak lortu ditugu.

Hitz gakoak: fraseologia konputazionala, idiomatikotasuna, esapide idiomatikoak, kolokazioak.

Abstract: The goal of this research is to develop and experimentally test different techniques for the automatic extraction of phraseological units (PUs) of noun+verb structure in Basque and for their characterization according to the idiomaticity level. Idiomaticity is considered the defining feature of the concept of phraseological unit (PU), ande we have measured its following components: institutionalization (statistical idiosyncrasy), semantic non-compositionality, morphosyntactic fixedness and lexical fixedness. The results show that the standard cooccurence techniques are significantly ourtperformed by semantic measures, and, to a lower extent, by measures of morphosyntactic flexibility. The results of lexical flexibility are poorer than expected. Finally, we obtain experimental evidence for several predictions of phraseological theory.

Keywords: computational phraseology, idiomaticity, idioms, collocations.

1. **SARRERA**

Jakina da konbinazio-sistema diskretua izatea dela hizkuntzaren ezaugarri gakoetako bat [1]. Hau da, lehendik inoiz sortu gabeko konbinazio berriak era ditzakegu multzo mugatua osatzen duten elementu bakunak konbinatuz, eta horrexetan datza hizkuntzaren adierazte-ahalmena [2]. Baina gaur egun aski onartua dago hizkuntzaren funtzionamendua ezin dela osagai bakunen konbinazio libreaz soilik azaldu [3], ikerketek erakutsi baitute hiztunok nolabaiteko «unitate aurrez eratu» batzuk erabiltzen ditugula, unean-unean egindako konbinazio libreen gisa berean eratzen ez direnak.

Horrelako konbinazioei *hitz anitzeko unitate* edo *unitate fraseologiko* (UF) deritze. Horrelakoak dira, esaterako, *adarra jo* eta *zarata atera*, *liburua irakurri* eta antzeko konbinazio libreek ez bezalako ezaugarriak dituztenak. Esanahiari soilik erreparatuz, lehenarena ezin ondoriozta daiteke osagaien esanahietatik; bigarrenean, *atera* aditzak adiera berezia du, gertuago baitago 'egin' edo 'sortu' adieretatik, *atera* ren ohiko adieratik baino.

UFek leku nabarmena dute hizkuntzaren fenomenoa esplikatu nahi duten teorietan, hizkuntzari buruzko informazioa biltzen duten hizkuntza-baliabideetan eta, azken urteotan argi ikusi denez, hizkuntzaren prozesamendu automatikoan (HP) [4].

Hizkuntza garatu eta erabili ahala, UF berriak agertzen dira testuetan, eta, horiek eskuratu eta prozesatu gabe, hizkuntza modernoaren prozesamenduak hutsuneak ditu. Esaterako, autore batzuek ohartarazi dute komunikabideetan kolokazio berriak erabiltzen direla [5], eta orobat esan liteke testu espezializatuez [6]. UFak ez dira multzo itxi bat, beraz.

Euskararen kasuan, hitz anitzeko unitateen prozesamenduan egindako aurrerapausoak terminologiaren erauzketara bideratu dira [7, 8, 9] eta corpusean automatikoki etiketatzera datu-base lexikaletan bilduta dauden UFak [10]. Ondorioz, euskararen prozesamendu automatikoak ez du ahal bezainbeste jorratu fraseologiaren arloa, eta oraindik onura atera dezake UFen erauzketa eta karakterizazio automatikotik.

Artikulu honetan laburtzen den tesi-lanaren bidez [11], ekarpen bat egin nahi izan dugu euskarazko fraseologia konputazionalaren arloan. Zehazki, euskarazko izena+aditza osaerako UFak corpusetik automatikoki erauzi eta idiomatikotasun-mailaren arabera karakterizatzeko lan esperimentala egin dugu.

2. **IDIOMATIKOTASUNA**

Idiomatikotasun terminoa erabat baliokide ez diren kontzeptuak adierazteko erabili da fraseologian.

Batetik, ez-konposizionaltasun semantikoarekin identifikatu da. Horren definizio modura, esan ohi da konbinazioaren esanahia ez dela osagaien esanahien konbinazioa edo batura, eta, beraz, ezin dela horien esanahiak konbinatuz lortu edo ulertu [12]. Hau da fraseologian tradizio handiena duen kontzepzioa, eta, neurri batean, oraindik ere onartuena dena. [13, 14, 15].

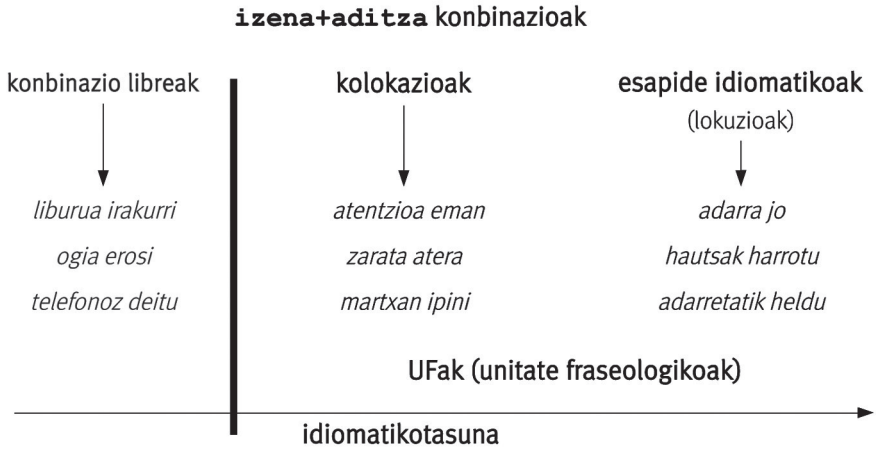
Bestetik, ikuspegi hori zabaltzen hasi da zenbait adituk nabarmendu dutenetik lokuzio edo esapide idiomatikoen propietateak ez direla esklusiboki semantikoak [16, 17], propietate gradualak direla [14], eta hitz anitzeko beste unitate mota batzuetan ere aurkitzen direla [18].

Ikuspegi horrek indar hartu du azken urteotan, batez ere fraseologia konputazionalean [10]. Horren arabera, hauek dira idiomatikotasunaren osagaiak:

— Instituzionalizazioa (idiosinkrasia estatistikoa). Hiztunek unitatetzat hautematen dute konbinazioa. Osagaiek halako joera bat dute elkarrekin konbinatzeko, eta emaitza «ohikoa» edo «ezaguna» da.
Estatistikoki idiosinkratikoa izatea ez dago zehatz-mehatz lotuta konbinazioaren maiztasunarekin, maiztasun hori osagaiak zori hutsez konbinatuko balira espero litekeena baino handiagoa izatearekin baizik [19, 15].
— Ez-konposizionaltasun semantikoa: konbinazioaren esanahia ez da osagaien esanahien konbinazioa edo batura. *Liburua irakurri* konposizionala da (‘liburua irakurri’ = ‘liburua’+‘irakurri’), baina *adarra jo* ez (‘adarra jo’ ≠ ‘adarra’ + ‘jo’). Propietate bitartzat hartu izan da, baina azkenaldian ugariagoak dira konposizionaltasun-mailaz edo konposizionaltasun partzialaz mintzo direnak [20].
— Finkapena: oso maiz azpimarratzen da UFek ez dutela egitura bereko konbinazio libreen portaera lexiko-sintaktiko bera, zeren, horiek onartzen dituzten aldakuntzen aldean, zenbait murriztapen agertzen baitituzte [14, 15, 21]. Bi mota bereizi ohi dira:

  • Finkapen morfosintaktikoa: konbinazioa ezin da egitura morfosintaktiko guztietan erabili. Esaterako, *adarra jo* esapidean, *adar* izena ez da pluralean erabiltzen; ezin zaio determinatzailerik, adjektiborik eta bestelako modifikatzailerik erantsi (**adar bat jo nion*, **adar ederra jo zenion*); eta ez da perpaus erlatiboetan erabiltzen (**aurrekoan jo zenidan adarra ez zitzaidan gustatu*).
  • Finkapen lexikala: konbinazioaren osagaiak ezin dira sinonimoz edo erlazio semantiko estua duten hitzez ordezkatu, emaitza ez delako erabiltzen, edo ez dituelako jatorrizko konbinazioaren propietate idiomatiko berberak [21]. Esaterako: *ziria sartu*, baina *zotza sartu*?, *ezkarda sartu*?; gastuak murrizteaz mintzatuz, *gerrikoa estutu* diogu, baina *uhala estutu*?

Aurreko propietateak hein desberdinean konbinatzen dira, eta UFak sailkatzeko irizpideak ezartzearen emaitza continuum bat izaten da, argi bereizten diren kategoria diskretu sorta baino gehiago [22, 23]. Nolanahi ere, continuum horretan «eremu» edo zona desberdinak bereizteko ahaleginak egin izan dira [24], 1. irudian ageri den eran.



1. irudia. *izena+aditza* osaerako sintagma-unitateen idiomatikotasunaren continuuma.

Ezkerraldean, konbinazio libreak daude, UF ez direnak. Eskuinaldean, esapide idiomatikoak (lokuzioak). Konbinazio ez-konposizionalak dira (opakoak zein figuratiboak), eta lexikalki zein morfosintaktikoki ez dira oso malguak. Erdialdean, berriz, kolokazioak kokatu ohi dira [25]. Normalean, esan ohi da horiek erdikonposizionalak, lexikalki murriztuak eta morfosintaktikoki aski malguak direla esatea.

3. IDIOMATIKOTASUNAREN KARAKTERIZAZIO AUTOMATIKOA

3.1. UF hautagaien erauzketa

Karakterizazioa egin aurretik, ezinbestez erauzi behar dira hautagaiak. Horretarako, teknika linguistikoak erabili ohi dira. Aukera sinpleena prozesamendu linguistikorik gabeko testu gordina (*raw text*) erabiltzea da, baina prozedura horrek muga argiak ditu [26], eta aski onartua da prozesamendu linguistiko minimo bat behar dela [27].

Oro har, esan daiteke lematizazioa eta etiketatze morfosintaktikoa direla minimo hori adierazten duten estandarrak [25]. Azken urteetan, zen-

bait ikertzailek aldarrikatu dute analisi sintaktikoaren bidez hobetu daitekeela erauzketaren eraginkortasuna [28, 29].

3.2. **UF hautagaien karakterizazioa**

UFen karakterizazioa egiteko teknikak idiomatikotasunaren propietateen neurketan oinarritzen dira. Horretarako, aipatu propietateekin lotutako fenomeno edo behagaiak kuantifikatu ohi dira. Lau propietateetako bakoitza neurtzeko, behagai hauek erabili ohi dira, hurrenez hurren:

— Agerkidetza: konbinazioaren estatistikak osagaiak ausaz konbinatuko balira espero litezkeen estatistikekin konparatzea, elkartze-neurrien bidez [30].
— Antzekotasun distribuzionala: UFen karakterizaziora egokituta, honela formula daiteke Z. Harrisen hipotesia[1]: testuingurua eta haren osagaien testuinguruak zenbat eta desberdinagoak izan, UFa ez-konposizionala izateko aukera handiagoa da. Beraz, konbinazioaren testuinguruak osagaien testuinguruekin konparatzea da teknikaren oinarria. Testuingurua errepresentatzeko erabiltzen den ohiko eredua IR (informazio-berreskuratzea) arlorako garatu zen *Vector Space Model* (VSM) edo *bektore-espazioaren eredua* izenekoaren egokitzapena da, eta *Word Space Model* (WSM) edo *hitz-espazioaren eredua* izena ere eman ohi zaio [32]. Errepresentazio horiek antzekotasun-neurrien bidez konparatzen dira [33, 34].
— Malgutasun morfosintaktikoa neurtzen da, konbinazioaren portaera morfosintaktikoaren eta erreferentzia-portaera baten arteko distantziaren bidez. Bi erreferentzia erabil daitezke: a) portaera orokorra, hau da, kategoria-osaera bereko konbinazioen batez besteko portaera; eta b) osagaien portaera, hau da, konbinazioaren osagai batek beste osagaiaren kategoriako edozein hitzekin osatutako konbinazioen batez besteko portaera[2]. Portaerak aldakuntza morfosintaktikoen banaketak dira, eta banaketen arteko distantzia-neurriak erabili ohi dira portaerak konparatzeko [35, 36, 23, 37].
— Malgutasun lexikala neurtzen da, osagaien ordezkagarritasunaren bidez: konbinazioaren estatistikak osagaiak sinonimoz edo haiekin erlazio semantikoa duten hitzez ordezkatuz sortzen diren konbinazioen estatistikekin konparatzea. Aurrekoan bezala, banaketen arteko distantzia-neurriak erabili ohi dira [38, 39, 37].

---

[1] «Difference of meaning correlates with difference of distribution.» [31]

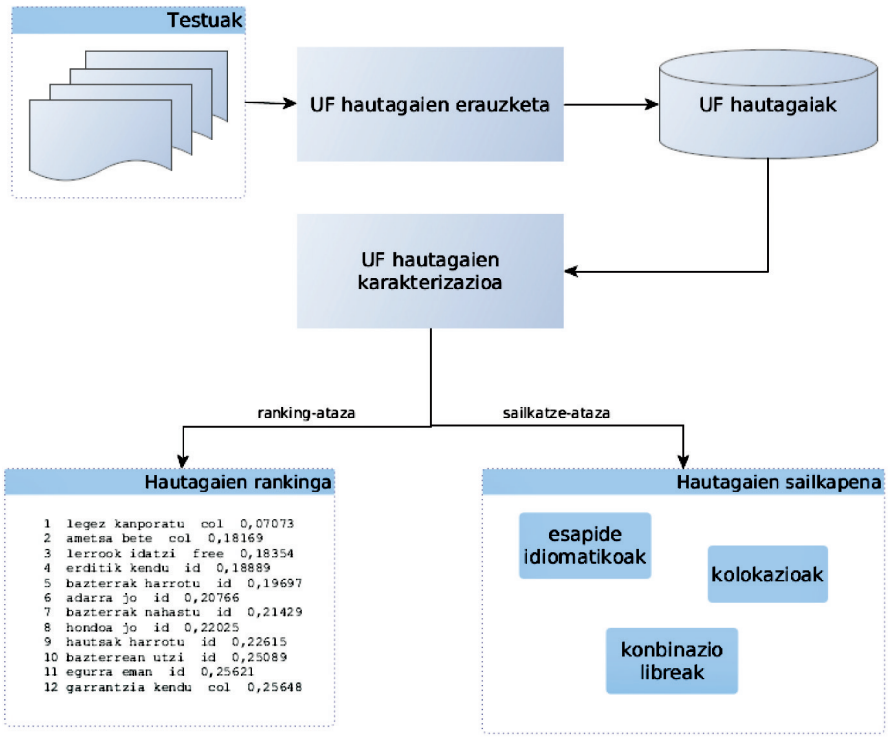
[2] Esaterako, *liburua irakurri* konbinazioaren portaera hauekin konparatzen da: *liburua* +aditza konbinazio guztien portaerarekin eta izena+*irakurri* konbinazioen portaerarekin.

Konparazio horien emaitza zenbat eta desberdinago, hautagaia hainbat eta idiomatikoago. Agerkidetza da gehien ustiatu den behagaia, eta hura neurtzeko elkartze-neurriak hasieratik erabili dira arlo honetako ikerkuntzan eta kolokazioak erauzteko tresnetan [40, 41]. Beste propietateak kuantifikatzeko ikerlanak nabarmen ugaritu dira azken hamarkadan [42, 39, 43].

4. LAN ESPERIMENTALAREN DISEINUA

Ikerkuntza-lan honen bidez, ikertze-galdera hauei erantzun nahi izan diegu:

— Zenbaterainoko korrelazioa dago idiomatikotasun-mailaren eta haren propietate bakoitzaren neurketen artean?
— Zenbateraino datoz bat UFen propietateen ebidentzia enpirikoak teoria fraseologikoak UFetarako zein UF-kategoria bakoitzerako aurresaten duenarekin?
— Hobetu daiteke UFen karakterizazioa idiomatikotasunaren propietate bakunen kuantifikazioaren emaitzak konbinatuz?



2. irudia. UFen erauzketa eta karakterizazio-atazak.

2. irudian eman dugu ikergai horien inguruko ebidentziak lortzeko diseinu esperimentalaren eskema orokorra. Bi urrats nagusi daude, hautagaien erauzketa eta ondorengo karakterizazio-atazak. Horiez gaien, aztergaitzat hartu ditugun *izena+aditza* konbinazioen egitura sorta eta karakterizazioa ebaluatzeko metodologia azalduko ditugu hurrengo ataletan.

2. irudian eman dugu ikergai horien inguruko ebidentziak lortzeko diseinu esperimentalaren eskema orokorra. Bi urrats nagusi daude, hautagaien erauzketa eta ondorengo karakterizazio-atazak. Horiez gaien, aztergaitzat hartu ditugun *izena+aditza* konbinazioen egitura sorta eta karakterizazioa ebaluatzeko metodologia azalduko ditugu hurrengo ataletan.

4.1. **UF hautagaiak erauztea**

Esperimentuetarako, 74 milioi hitzeko kazetaritza-corpus bat erabili dugu, *Euskaldunon Egunkari*ko eta *Berria* ko artikuluez osatua.

UF hautagaiak testutik automatikoki erauzteko eta forma kanonikoa esleitzeko prozesu automatikoa garatu da. Erauzketa-prozesuaren deskribapen zehatza [44] lanean egin da. Bi urrats ditu: bigrama-sorkuntza eta bigramen forma kanonikoa lortzeko normalizazioa[3]. Lehen urratserako, Ngram Statistics Package erabili dugu[4]. Erauzketa egin aurretik, corpusa linguistikoki prozesatu da, Ixa taldearen Eustagger etiketatzailearen bidez lehenik [45]; ondoren, ikerkuntza honetan landutako prozesu batzuk inplementatu dira, corpusak erauzketarako behar diren ezaugarriak izan ditzan.

4.2. **Karakterizazio-atazak**

Bi karakterizazio-ataza bereizi ditugu: a) ranking-ataza, zeinetan, UFen karakterizaziorako azaldu ditugun teknikak erabiliz, propietate bakoitzaren neurketen emaitzak hautagaiak ordenatzeko erabili baitira; eta b) sailkatze-ataza, zeinetan esperimentu bakunen emaitzak konbinatu baititugu, ikasketa automatikoko teknikak erabiliz.

4.3. ***izena+aditza* konbinazio aztergaiak**

1. taulan eman dugu aztergai izan ditugun izena+aditza osaerako konbinazio-moten errepertorioa.

---

[3] Normalizazioaren bidez, forma kanoniko bakarrean konputatzen dira *egunkaria*/*egunkariak*/*egunkariok*/*egunkari*/*egunkaririk* formek *irakurri* aditz-lemarekin osatutako bigramak.

[4] http://search.cpan.org/dist/Text-NSP. Kontsulta-data 2015/05/24.

1. taula. Ikerketa honetako 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, aukera eman, erabakia hartu, zarata atera, 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$_{\text{PS*}}$ + aditza: *mendean hartu, martxan jarri, adarretatik heldu, larrutik ordaindu, burutik kendu, harira etorri, gogora ekarri, aurrera eraman, sareetara bidali, muturreraino eraman, aurrez ikusi, oinarritzat hartu*

* PS: postposizio-sintagma.

4.4. **Ebaluazioa: metodologia eta baliabideak**

Ranking-atazan, heinen korrelazio-koefizienteak erabili ohi dira, eta guk Kendall τ aukeratu dugu, berdinketak kudeatzeko aldaeran (τ_B). Batez besteko doitasuna ere (*AP-Average Precision*) erabili dugu. Lehenik, AP_{UF}, esapide idiomatikoen (id) eta kolokazioen (col) arteko bereizketa kontuan hartzen ez duena; bestetik, *AP* espezifikoak, AP_{id} eta AP_{col}. Sailkapen automatikoaren bidezko atazen ebaluazioan, Weka tresnak eskaintzen dituen zenbait neurri erabili ditugu[5] [46]: ondo sailkatutako instantzia kopurua (*CC*), klase bakoitzaren *F* neurriak, F_{mikro} eta F_{makro}.

Ebaluazio-erreferentzia edo *gold standard* tzat, bigrama-erauzketa batetik ausaz ateratako eta aditu-talde batek eskuz sailkatutako 1 145 konbinazioko multzo bat erabili dugu, hiru kategoriatan banatua: esapide idiomatikoak (80), kolokazioak (268) eta konbinazio libreak (797). Anotatzaileen artean, 0,55eko Fleiss $\varkappa$ lortu dugu; adostasun-maila ertaina da [47].

[5] http://www.cs.waikato.ac.nz/ml/weka/. Kontsulta-data 2015/05/24.

Horren aurretik, ordea, aipatutako bigrama-erauzketaren parametroak aukeratu behar izan ditugu. Esperimentu batzuk egin ditugu leiho-zabaleraren, maiztasun minimoaren eta bigrama-normalizazioaren eragina neurtzeko. Horiek ebaluatzeko, geuk sortutako hiztegi-erreferentzia bat erabili dugu, iturri hauetako *izena + aditza* unitateez osatua: a) Euskaltzaindiaren *Hiztegi Batua*[6]; b) Ibon Sarasolaren *Euskal Hiztegia* [48]; c) Elhuyar Fundazioaren *Euskara-Gaztelania /Castellano-Vasco Hiztegia* [49]; d) *Intza proiektua*[7]; e) Ixa taldearen EDBL datu-base lexikala[8] [50]. Esperimentu horien emaitzetatik ondorioztatu dugu hauek direla ebaluazio-erreferentziaren oinarria izango den erauzketa lortzeko parametro egokienak: $w = \pm 1$ leiho-tamaina, $f \geq 30$ eta bigrama normalizatuak.

Hiztegi-erreferentziak badu beste alderdi interesgarri bat: eskuz landutako ebaluazio-erreferentziarekin konpara daiteke, bi erreferentzien osaerari buruzko datuak ateratzeko.

2. taulan daude ebaluazio-erreferentziaren eta hiztegi-erreferentziaren arteko konparazioaren emaitzak. Lehen ondorio modura, esan dezakegu honelako lan batek aukera ematen duela hiztegia aberasteko. Batez ere, kolokazioen kategorian da ondorio hori nabarmena, hiztegian daudenak % 15,3 baino ez baitira.

5. IDIOMATIKOTASUNA KARAKTERIZATZEKO ESPERIMENTUAK

4.2. atalean aurkeztutako bi karakterizazio-atazetan egindako esperimentuak deskribatuko ditugu jarraian.

2. taula. Eskuz sailkatutako ebaluazio-erreferentziaren eta hiztegi-erreferentziaren arteko konparazioa.

Kategoria	Hiztegi-erreferentzia		
	Bai	Ez	Totala
id	23	57	80
col	41	227	268
free	10	787	797
Totala	74	1.071	1.145

[6] http://www.euskaltzaindia.net/hiztegibatua. Kontsulta-data: 2010-06-04.
[7] http://intza.armiarma.com. Kontsulta-data: 2010-06-02.
[8] http://ixa2.si.ehu.es/edbl. Kontsulta-data: 2010-06-22.

5.1. **Propietateen banakako neurketa**

5.1.1. *Idiosinkrasia estatistikoa, agerkidetza neurtuz*

Bigramen agerkidetza-datuen analisi estatistikoa S. Everten UCS toolkit[9] paketearen bidez egin dugu [30]. Elkartze-neurri hauek kalkulatu ditugu: *z* neurria, *t* neurria, khi karratua (χ^2), egiantz-arrazoiaren logaritmoa, Fisherren test zehatza, elkarrekiko informazioa (MI), MI^3 eta *f*.

5.1.2. *Konposizionaltasun semantikoa, antzekotasun distribuzionala neurtuz*

Hauek dira UFen konposizionaltasuna karakterizatzeko gure metodologiaren oinarriak:

— UF hautagai bakoitzaren testuinguruak haren osagai bakunen testuinguruekin konparatzea. Konparazio hori osagai bakoitzarekiko egin dugu, eta 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 hartu. Horrela jokatuz, ahal den gehiena bereizi nahi izan dugu konbinazioaren eragina osagaien testuinguruen modelizazioan, eta, horretara, konparazioa adierazgarriagoa da.
— Testuinguruak sortzeko zenbait irizpide:

- Testuinguru-dokumentuak elikatzeko, eduki-hitzak bakarrik erabiltzea (izenak, aditzak eta adjektiboak).
- Esaldi osoa hartu dugu testuingurutzat, [51] lanari jarraituz.
- Ondoz ondoko agerkidetzak hartu dira bigramatzat; izan ere, komeni da erauzketan ezarritako irizpide bera erabiltzea ($w = \pm 1$).
- Izenen kasuan, lemaz gain, kasua ere hartu da kontuan[10]; gainerakoetan, lema hutsa.

Antzekotasun distribuzionaleko tekniken bidez egin dugu testuinguruen konparazioa [52].

---

[9] http://www.collocations.de/software.html. Kontsulta-data: 2010/09/12.

[10] Esaterako, esaldi bat *egunkarian irakurri* bigramaren testuingurutzat hartzeko, ez da aski *egunkari* izena eta *irakurri* aditza agerkide izatea, *egunkari* lemaren kasua inesiboa izatea ere behar da.

Zehazki, teknika hauek erabili ditugu:

— WSM (*Word Space Model*) ereduan ohikoak diren antzekotasun-neurrietatik, Jaccard koefizientea, kosinua eta Jensen-Shannon dibergentzia aukeratu ditugu.
— Berry-Rogheren *R* balioa [53] eta Wullfek egindako bi hedapenak [23].
— VSMren inplementazio berezia den ezkutuko semantikaren analisiaren (LSA, *Latent Semantic Analysis*) aplikazioa aztertu dugu. Infomap softwarea erabili dugu horretarako[11].

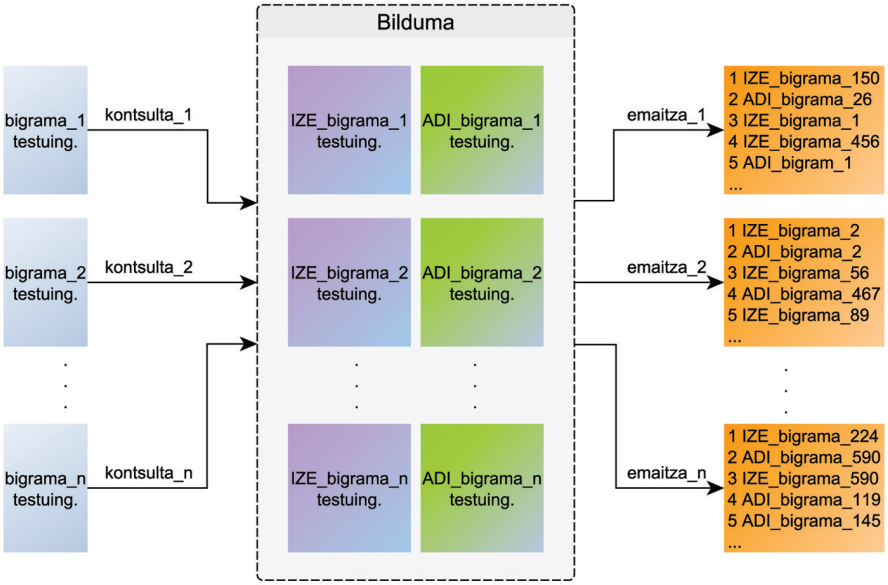
Bestetik, UFen konposizionaltasuna neurtzeko argitaratu diren ikerlanetan IR sistemez baliatzen den lanik aurkitu ez badugu ere, komenigarritzat jo dugu dokumentu arteko antza kalkulatzeko erabili ohi diren indize batzuk ere aplikatzea. Lemur Toolkit[12] [54] aukeratu dugu horretarako. Lemurrek dakartzan indize hauek erabili ditugu: Indri, *tf-idf*, KL dibergentzia eta Okapi. Gainera, Lemurrek dakarren kosinuaren inplementazioa ere erabili dugu[13].

Funtsezko prozedura hau da: bigramen testuinguruen dokumentuak *query* edo kontsulta gisa erabiltzen dira bigramen osagaien testuinguruen dokumentuek osatzen duten bildumaren kontra. Horretara, uneko bigramaren testuinguru-dokumentuen antzekoenak diren testuinguru-dokumentuak lortzen ditugu. Ideia hori bi eratara inplementatu dugu:

— L1: bektoreak osatzeko egin dugunaren antzera, bigrama baten testuinguruak dokumentu bakarrean bildu dira, eta osagaien testuinguruak dokumentu banatan. 3. irudian ageri denez, bigrama bakoitzetik hiru dokumentu lortu dira: batetik, bigramaren testuinguruak dituena, kontsulta gisa erabilia; eta, bestetik, 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 sortu dira. Osagaien dokumentuak, berriz, L1 esperimentuan bezala tratatu dira.

3. irudian dugu L1 modalitatearen eskema.

[11] http://infomap-nlp.sourceforge.net/. Kontsulta-data 2013/06/03.
[12] http://www.lemurproject.org. Kontsulta-data 2013/12/19.
[13] http://www.lemurproject.org/doxygen/lemur/html/RetEval.html. Kontsulta-data 2010/06/15.



3. irudia. Lemurrekin egindako L1 modalitateko kontsultak eta emaitzak.

5.1.3. *Malgutasun morfosintaktikoa, erreferentzia-portaerarekiko distantzia neurtuz*

Metodologiaren oinarria konparazio bat da: aztergai dugun bigrama bakoitzaren portaera morfosintaktikoa erreferentzia-portaera batekin konparatzea. Horretarako, bi konparazio-prozedura hauek probatu ditugu:

— Portaera orokorrarekiko konparazioa [35, 43, 23]: kategoria-osaera bereko (gurean, *izena + aditza*) konbinazioen batez besteko portaerarekikoa.
— Osagaien portaerarekiko konparazioa [36]: konbinazioaren osagai batek beste osagaiaren kategoriako edozein hitzekin osatutako konbinazioen batez besteko portaerarekikoa.

Kontuan hartutako aldakuntzak. Malgutasun morfosintaktikoa neurtzeko aztertzen diren aldakuntzen multzoa hizkuntzaren ezaugarrien araberakoa da. Arlo honetako ikerlan esanguratsu batzuk [55, 56] eta EDBLko aditz-lokuzioen gauzatze-eskemak [10] kontuan izanik, kasu hauek hautatu ditugu malgutasun morfosintaktikoa karakterizatzeko:

— Izenaren ezker- eta eskuin-hedapenak: determinatzailea (*liburu **bat** irakurri dut*); ize-nondoa (*liburu **interesgarria** irakurri nuen*); izen-

laguna (***gustuko*** *liburuak irakurtzea*); eta erlatiboa (***irakurri dudan*** *liburua; anaiak* ***irakurritako*** *liburu batzuk*).
Hedapen bat baino gehiago konbina daitezke aldakuntza berean: *liburu interesgarri bat irakurri dut; lau liburu hauek irakurri ditut, anaiak irakurritako frantsesezko liburu eder batzuk*.

— Mugatasuna: izenaren eta 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*.
— Osagaien ordena (IZE ADI / ADI IZE): *liburua irakurri dut /irakurri dut liburua*.

Azaleko sintaxiko murriztapen-gramatika bat garatu dugu, eta portaera bakoitza detektatzeko egokitu dugu.

Bigrama baten portaera zein bi konparazio-prozeduretako portaerak ezagutzeko, konbinazio bakoitzaren aldakuntzen maiztasunak kontatzen ditugu.

Kontaketa horien emaitzak aldakuntza morfosintaktikoen banaketak dira, distantzia-neurrien bidez konparatzen ditugunak.

Neurri hauek erabili ditugu portaeren arteko distantzia kalkulatzeko: Kullback-Leibler dibergentzia [43]; Wullfen SSD neurria (*sum of squared deviations*) eta entropia erlatiboa [23]; eta Bannarden CPMI (*conditional pointwise mutual information*) [36].

5.1.4. *Malgutasun lexikala, osagaien ordezkagarritasuna neurtuz*

Bigramen osagaien ordezkagarritasuna neurtzeko, osagaien ordezkoak behar dira, eta baliabide hauek erabili ditugu horiek lortzeko: Elhuyarren *Sinonimoen Kutxa*[14]; eta Ixa taldearen Euskal WordNet [57]. Halaber, baliabideon estaldura handitzeko, konbinatu egin ditugu, baita hedatu ere, bigramaren osagai bakoitzaren Euskal WordNeteko senideak (*siblings*) gehituz. Azkenik, [39]-ri jarraituz, bigramen osagaien thesaurus distribuzional bat osatu dugu corpusetik, semantikoki antzekoenak diren ordezkoak eskuratzeko (kategoria berekoak, betiere). Lan hori egiteko, [9] lanean garatutako tresna egokitu dugu.

Ordezkagarritasuna neurtzeko, R_{nv} eta R_{vn} indizeak [39], eta $\text{Fixedness}_{\text{lex}}$ neurria [43] erabili ditugu.

[14] *Sinonimoen Kutxa*. 2010. 3. ed. Elhuyar Fundazioa. Usurbil.

	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

4. irudia. *adostasuna lortu* bigramaren aldakuntza batzuk (izenaren ezker- eta eskuin-hedapenak), eta erauzten diren hedapenak.

5.1.5. *Esperimentu bakunen emaitzen analisia*

3. taulan laburbildu ditugu teknika bakoitzaren bidez lortutako emaitza onenak. 5. irudian daude emaitza horien *P* /*R* (doitasuna/estaldura) kurbak.

Emaitza horien alderdi esanguratsuenak:

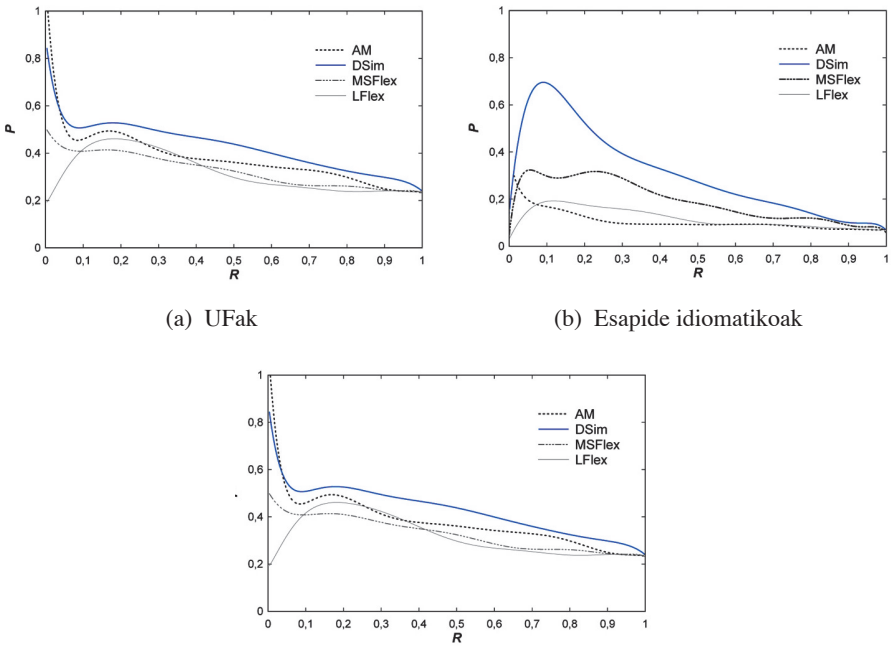
— Antzekotasun distribuzionaleko neurriak dira idiomatikotasun-mailarekin τ_B korrelazio onena dutenak. Zehazki, testuinguru-dokumentuen arteko antza neurtzen duten Indri indizea eta KL dibergentzia dira neurri onenak, L2 esperimentu-modalitatean. Bestetik, horien nagusitasuna are nabariagoa da esapide idiomatikoen erauzketan.

— Kolokazio-erauzketan, aditzarekiko antzekotasun distribuzionala neurtzen duen Indri indize batek izan du emaitza onena, kolokazioen ezaugarri aipatuenetakoa den idiosinkrasia estatistikoa neurtzen duten AMen gainetik (AM onena *t* neurria da). Emaitza hori oso garrantzitsua da, kolokazioak erdikonposizionalak direlako ideiarekin bat baitator.

3. taula. Esperimentu bakunetan teknika bakoitzaren bidez lortutako emaitza onenak (AM: *association measures* edo elkartze-neurriak; DSim: antzekotasun distribuzionala; MSFlex: malgutasun morfosintaktikoa; LFlex: malgutasun lexikala).

Teknika	τ_B	AP_{UF}	AP_{id}	AP_{col}
ausaz	0,000	0,309	0,070	0,234
AM	0,197	0,455	0,119	0,383
DSim	**0,322**	**0,566**	**0,320**	**0,431**
MSFlex	0,154	0,434	0,202	0,331
LFlex	0,110	0,381	0,122	0,323

— AMak: emaitza hobeak kolokazioetan esapide idiomatikoetan baino.
— MSFlex emaitza onenak esapide idiomatikoetan lortu dira (AP_{id}; zehazki, mugatasun-aldakuntzen neurketetan). Bigarren onena da, AMen gainetik. Gainerakoan, ez dituzte AMen emaitzak gainditzen.
— LFlex neurketetan, emaitzak espero baino txarragoak dira.



(a) UFak (b) Esapide idiomatikoak

(c) Kolokazioak

5. irudia. Idiomatikotasun-rankingen *P*/*R* (doitasuna/estaldura) kurbak.

5.2. **Propietateen integrazioa: sailkapen automatikoa**

Ikasketa automatikoko esperimentuak egiteko, Weka paketea erabili dugu. Sei algoritmorekin probak egin ondoren [58], bi hauekin lortu ditugu emaitza onenak[15]: *Logistic Regression* (LR) eta SMO (*Sequential Minimal Optimization*)[16].

Atributu gisa, aurreko ataleko neurketen emaitzak erabili ditugu, eta konbinazioaren aditza ere bai[17]. Ebaluazioan, balidazio gurutzatua erabili dugu, ebaluazio-erreferentzia ez delako oso handia (1 145 instantzia). Atributuak hautatzeko iragazkiak ikaste-multzoan dauden instantziak soilik aztertzen ditzan, AttributeSelectedClassifier metasailkatzailea hautatu dugu.

Sistema hori erabiltzen denean, ordea, kontuz ibili behar da atributuak automatikoki hautatzeko iragazkiekin, oso garrantzitsua baita iragazkiak ikaste-multzoan dauden instantziak soilik aztertzea, test-multzotik independenteki. Wekak atributu-iragazketa automatikoa baldintzapen horietan ondo egiteko eskaintzen dituen metodoen artean, *AttributeSelectedClassifier* metasailkatzailea hautatu dugu[18].

4. taulan, lau datu-multzo hauetan lortutako emaitzak bistaratu ditugu: 1) DSim (antzekotasun distribuzionaleko neurketak)[19]; 2) 4 osag. (lau propietateen neurketak); 3) 4 osag.+ad. (aurreko atributuak eta aditza); 4) CSBF (AttributeSelectedClassifier meta-sailkatzailean, CfsSubsetEval1 ebaluatzaileaz, eta BestFirst bilaketa-metodoaz osatutako iragazkia).

Emaitza onenak SMO algoritmoarekin lortu dira, 4 osag.+ad. datu-multzoa erabiliz. Jakina da SVM familiako algoritmoek hobeto kudeatzen dutela atributu asko erabiltzeak sortu ohi duen zarata. Aditza kontuan hartzeak eragin positiboa du esapide idiomatikoen eta kolokazioen sailkapenean, baina ez du eraginik konbinazio libreenean.

Hala ere, datu-multzo hori osatzeko, prozesamendu handia behar da, eta LR metodoak atributu-hautaketa automatikoaren bidez lortutako emaitzek aukera eman lezakete bideragarritasunaren eta emaitzen kalitatearen arteko erlazio on bat lortzeko. Automatikoki hautatuko atributu gehienak DSim propietatekoak dira (16), baina MSFlex eta aditz motako atributuek

---

[15] Hona gainerako lauak: Naive Bayes, C4.5 decision tree (j48), RandomForest eta PART.

[16] *Support Vector Machine* (SVM) edo sostengu-bektoreen makinen modalitatearen inplementazio bat.

[17] Aditza atribututzat hartu dugu [43] lanari jarraituz, eta gure erreferentzian ere aditzaren eta UF motaren arteko korrelazio esanguratsua dagoela egiaztatu ondoren.

[18] http://weka.wikispaces.com/Performing+attribute+selection. Kontsulta-data 2015/05/24.

[19] Beste hiru propietateen neurketen datu-multzoak ere probatu ditugu, baina, espero izatekoa zenez, emaitzak DSim-enak baino txarragoak izan dira.

4. taula. Ikasketa automatikoko esperimentuen emaitzak.

Atrib.	Metod.	*CC*	F_{id}	F_{col}	F_{free}	F_{mikro}	F_{makro}
Oinarri-lerroa		69,607	0,000	0,000	0,821	0,571	0,274
DSim	LR	74,061	0,270	0,468	0,842	0,714	0,527
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
4 osag.+ad.	SMO	**76,856**	**0,418**	**0,544**	**0,858**	**0,754**	**0,607**
CS-BF	LR	75,721	0,339	0,487	0,854	0,732	0,560

ere laguntzen dute emaitzak hobetzen (hurrenez hurren, 7 eta 9); ekarpen txikiena AM atributuek (3) eta LFlex motakoek egiten dute (2).

Beraz, konposizionaltasun-maila da idiomatikotasuna karakterizatzeko propietate eraginkorrena, baina beste osagai guztiek ere bere ekarpena egiten dute.

6. **ONDORIOAK ETA ETORKIZUNEKO LANAK**

Ondorio nagusi modura, esan dezakegu arlo honetan estandar diren agerkidetza-tekniken emaitzak modu esanguratsuan gainditu direla, batez ere teknika semantikoen bidez, baina baita malgutasun morfosintaktikoaren neurketaren bidez ere. Aldiz, finkapen lexikalaren neurketak emaitza txarragoak izan ditu teoria fraseologikoaren aurresanak eta lan esperimental batzuen emaitzak kontuan izanik espero zitezkeenak baino. Ikasketa automatikoan, ikusi dugu UF-kategoriaren eta konbinazioaren aditzaren arteko korrelazioa erabilgarria dela.

Bestetik, fraseologiaren aurresan batzuen ebidentzia esperimentalak lortu ditugu: a) idiomatikotasunaren izaera konplexua; b) idiosinkrasia semantikoaren nabarmentasuna, eta konposizionaltasunaren gradualtasuna; eta c) kolokazioen erdikonposizionaltasuna eta malgutasun morfosintaktiko handia.

Ikerketa honen ekarpenak baliagarriak dira etorkizuneko hiztegigintzak automatizaziorantz izango duen bilakabidean[20], eta hizkuntzaren prozesa-

[20] Esaterako, 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: http://webcorpusak.elhuyar.org/cgi-bin/kolokatuak.py. Kontsulta-data 2015/05/24.

menduko arloko zenbait atazatan, hala nola datu-base lexikalen elikatzean, corpusen etiketatzean eta, testuinguru eleaniztunean aplikatuta, itzulpen automatikoan.

Etorkizuneko lanei dagokienez, egiaztatu behar litzateke erauzketa hobetzen den informazio sintaktiko aberatsagoa erabiliz (adib., osagaien mendekotasunak). Bestetik, interesgarria da interpretazio literala eta idiomatikoa izan ditzaketen konbinazioen agerpenak bereiztea. Azkenik, bi hizkuntzatako UF bikote baliokideak erauzteko eta karakterizatzeko, komeni da garatutako metodologia corpus paraleloei aplikatzea.

Aipamenak

Tesi-lan hau Elhuyar Fundazioaren KONBITZ eta KONBITZ2 proiektuen testuinguruan egin da (EJren Ekonomia Garapena eta Lehiakortasuna sailaren SAIOTEK 2011 eta SAIOTEK 2012 programak).

Elhuyarko I+Gko eta Ixa taldeko hainbat adituk lagundu digute ikerketan honetan. Eskerrik asko denoi.

7. **BIBLIOGRAFIA**

[1] PINKER, S. 1994. *The Language Instinct*. New York: William Morrow and Company.

[2] 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.

[3] FILLMORE, C.J. 1979. «On fluency». *Individual Differences in Language Ability and Language Behavior*, D. KEMPLER eta W.S.Y. WANG, ed. 85-101. Academic Press, New York.

[4] 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.

[5] ALTZIBAR, X. 2005. «Kolokazioak euskaraz. Zer axola duten kazetaritzan». *Euskarazko kazetaritzaren I. kongresua. Kazetaritza euskaraz: oraina eta geroa,* 383-395. UPV/EHU.

[6] 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.

[7] ALEGRIA, I., GURRUTXAGA, A., LIZASO, P., SARALEGI, X., UGARTETXEA, S. eta URIZAR, R. 2004. «An Xml-Based Term Extraction Tool for Basque». *Proceedings of the 4th International Conference on Language Resources and Evaluations-LREC 2004*. Lisboa.

[8] 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.

[9] 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.

[10] URIZAR, R. 2012. *Euskal lokuzioen tratamendu konputazionala*. Doktorego-tesia, Euskal Filologia saila, UPV/EHU.

[11] GURRUTXAGA, A. 2014. *Idiomatikotasunaren katakterizazio automatikoa: izena+aditza konbinazioak*. Doktorego-tesia, Informatika Fakultatea, UPV/EHU, Donostia.

[12] ZULUAGA, A. 1980. *Introducción al estudio de las expresiones fijas*. Frankfurt: P.D. Lang.

[13] GLÄSER, R. 1998. «The stylistic potential of phraseological units in the light of genre analysis». *Phraseology: Theory, Analysis, and Applications*, A. Cowie, ed. 125-143. Oxford University Press, AEB.

[14] RUIZ GURILLO, L. 1998. «Una clasificación no discreta de las unidades fraseológicas del español». *Estudios de fraseología y fraseografía del español actual*, G. WOTJAK, ed. Lingüística Iberoamericana, 13-37.

[15] MANNING, C.D. eta SCHÜTZE, H. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

[16] FERNANDO, C. eta FLAVELL, R. 1981. *On Idiom: Critical Views and Perspectives*. University of Exeter.

[17] 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.

[18] COWIE, A.P., MACKIN, R. eta MCCAIG, I. 1983. *Oxford Dictionary of Current Idiomatic English*. Oxford University Press.

[19] GRIES, S.T. 2008. «Phraseology and linguistic theory: A brief survey». *Phraseology: An Interdisciplinary Perspective*, S. GRANGER eta F. MEUNIER, ed. 3-25. John Benjamins Publishing, Amsterdam/Filadelfia.

[20] MOON, R. 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Clarendon Press Oxford.

[21] CONTRERAS, J.M. eta SUÑER, A. 2004. «Los procesos de lexicalización». *Las fronteras de la composición en lenguas románicas y en vasco*, I. ZABALA, E. PÉREZ GAZTELU, eta L. GARCÍA, ed. 47-108. Universidad de Deusto, Servicio de Publicaciones, Donostia.

[22] SINCLAIR, J. 1996. «The search for units of meaning». *Textus* **9**, 1, 75-106.

[23] WULFF, S. 2008. *Rethinking Idiomaticity*. Corpus and Discourse. Continuum International Publishing Group Ltd, New York.

[24] COWIE, A. 1998. *Phraseology: Theory, Analysis, and Applications*. Oxford University Press, AEB.

[25] HEID, U. 2008. «Computational phraseology. An overview». *Phraseology: An Interdisciplinary Perspective*, S. Granger eta F. Meunier, ed. 337-360. John Benjamins Publishing, Amsterdam/Filadelfia.

[26] EVERT, S. 2008. «Corpora and collocations». *Corpus Linguistics. An International Handbook*, A. LÜDELING eta M. KYTÖ, ed. De Gruyter Mouton, **29**, 1212-1247.

[27] SERETAN, V. 2011. *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Springer, Dordrecht.

[28] 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.

[29] PECINA, P. 2009. *Lexical Association Measures: Collocation Extraction*, vol. 4 of *Studies in Computational and Theoretical Linguistics*. Institute of Formal and Applied Linguistics, Charles University, Praga, Txekia.

[30] EVERT, S. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Doktorego-tesia, University of Stuttgart.

[31] HARRIS, Z.S. 1970. *Papers in Structural and Transformational Linguistics*. D. Reidel, Dordrecht, Herbehereak.

[32] 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.

[33] WEEDS, J.E. 2003. *Measures and Applications of Lexical Distributional Similarity*. Doktorego-tesia, University of Sussex.

[34] KORKONTZELOS, I. 2010. *Unsupervised Learning of Multiword Expressions*. Doktorego-tesia, University of York.

[35] BARKEMA, H. 1994. «Determining the syntactic flexibility of idioms». *Realising and Using English Language Corpora,* 39-52.

[36] 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,* Association for Computational Linguistics, 1-8.

[37] FAZLY, A., COOK, P. eta STEVENSON, S. 2009. «Unsupervised type and token identification of idiomatic expressions». *Computational Linguistics* **35**, 1, 61-103.

[38] LIN, D. 1999. «Automatic identification of non-compositional phrases». *Proceedings of the 37th Annual Meeting of the ACL*, Association for Computational Linguistics, 317-324.

[39] VAN DE CRUYS, T. eta MOIRÓN, B. 2007. «Semantics-based multiword expression extraction». *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions,* Association for Computational Linguistics, 25-32.

[40] CHURCH, K. eta HANKS, P. 1990. «Word association norms, mutual information, and lexicography». *Computational Linguistics* **16**, 1, 22-29.
[41] SMADJA, F. 1993. «Retrieving collocations from text: Xtract». *Computational Linguistics* **19**, 1, 143-177.
[42] 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,* Association for Konputazional Linguistics, 89-96. Sapporo, Japonia.
[43] 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,* Association for Computational Linguistics, 9-16.
[44] 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. ACL HLT 2011,* 2-7. Portland, AEB.
[45] 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.
[46] 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.
[47] LANDIS, J.R. eta KOCH, G.G. 1977. «The measurement of observer agreement for categorical data». *Biometrics,* 159-174.
[48] SARASOLA, I. 1996. *Euskal Hiztegia*. Kutxa Fundazioa / Fundación Kutxa, Donostia.
[49] ELHUYAR. 2006. *Elhuyar Hiztegia. Euskara/Gaztelania-Castellano/Vasco*. Elhuyar Fundazioa, Usurbil.
[50] 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.
[51] 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,* Association for Computational Linguistics, 54-60.
[52] 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.

[53] BERRY-ROGGHE, G. 1974. «Automatic identification of phrasal verbs». *Computers in the Humanities*, 16-26.

[54] 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*.

[55] 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.

[56] ODRIOZOLA, J. 2010. «Euskararen aditz-unitate fraseologikoen deskribapena». Unibertsitateko katedra plazarako lehiaketa. Zientzia eta Teknologia fakultatea.

[57] POCIELLO, E., AGIRRE, E. eta ALDEZABAL, I. 2011. «Methodology and construction of the Basque WordNet». *Language Resources and Evaluation*, **45**, 2, 121-142.

[58] GURRUTXAGA, A., ALEGRIA, I. eta LARDIZABAL, M. 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*, Association for Computational Linguistics, 116-125. Atlanta, AEB.