I. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2015eko maiatzaren 13, 14 eta 15
Durango, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Euskarazko izena+aditza konbinazioak corpusetik automatikoki erauztea eta idiomatikotasunaren arabera karakterizatzea

A. Gurrutxaga, I. Alegria eta X. Artola

599-606 or.
https://dx.doi.org/10.26876/ikergazte.i.82

Euskarazko `izena+aditza` konbinazioak corpusetik automatikoki erauztea eta idiomatikotasunaren arabera karakterizatzea

Gurrutxaga, A.[1] **Alegria, I.**[2] eta **Artola, X.**[2]

[1] Elhuyar Fundazioa. Zelai Haundi kalea, 3, Osinalde industrialdea, 20170 Usurbil
[2] Ixa taldea - UPV/EHU. Manuel Lardizabal 1, 48014 Donostia

Laburpena

Euskarazko `izena+aditza` egiturako unitate fraseologikoak (UFak) corpusetik automatikoki erauzi eta idiomatikotasun-mailaren arabera karakterizatzeko lan esperimentala egin dugu. Corpusetik hautagaiak erauzteko sistema landu ondoren, idiomatikotasunaren lau ezaugarri edo propietateetako bakoitza neurtzeko teknikak garatu eta ebaluatu ditugu, hiru adituk eskuz sailkatutako erreferentzia erabiliz. Hiru kategoria bereizi dira: esapide idiomatikoa, kolokazioa eta konbinazio librea. Azkenik, ezaugarri bakunen neurketak ikasketa automatikoko sailkatze-ataza batean konbinatu dira. Ondorio nagusia da arlo honetan estandar diren agerkidetza-tekniken emaitzak modu esanguratsuan gainditu direla, batez ere teknika semantikoen bidez, baina baita malgutasun morfosintaktikoaren neurketaren bidez ere.

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

Abstract

We present an experimental study on the automatic extraction of phraseological units of noun+verb structure in Basque, and their characterization according to the idiomaticity level. After automatically extracting candidates from corpora, we develop several techniques for quantifying the four basic properties of idiomaticity, using for evaluation a gold standard of candidates classified by three experts. We use three categories: idioms, collocations and free combinations. Finally, the results of those experiments have been combined using Machine Learning for classification. The results show that the standard cooccurence techniques are significantly ourtperformed by semantic measures, and, to a lower extent, by measures of morphosyntactic flexibility.

Keywords: computational phraseology, idiomaticity, idioms, collocations

1 Sarrera eta motibazioa

Hitz anitzeko unitateek edo unitate fraseologikoek (UFek) egiteko giltzarria dute, hiztegigintzan eta hizkuntzen irakaskuntzan ez ezik, hizkuntzaren prozesamendu automatikoan ere (Sag *et al.*, 2010). Gaur egun aski onartua dago hizkuntzaren funtzionamendua ezin dela osagai bakunen konbinazio libreaz soilik azaldu, hiztunek erabiltzen dituzten elementu batzuk nolabaiteko unitate aurrez eratuak baitira (Fillmore, 1979, 92). Horrelakoak dira, esaterako, *adarra jo* eta *zarata atera*. Lehenaren esanahia ezin igarri osagaien esanahietatik; bigarrenean, *atera* aditzak adiera berezia du, gertuago baitago 'egin' edo 'sortu' adieratik, *atera*ren ohiko adieratik baino.

Euskararen kasuan, hitz anitzeko unitateen prozesamenduan egindako aurrerapausoak terminologiaren erauzketara bideratu dira (Alegria *et al.*, 2004, 2006; Saralegi *et al.*, 2008) eta corpusean automatikoki etiketatzera datu-base lexikaletan adierazita dauden UFak (Urizar, 2012).

Ikerkuntza honen bidez, ekarpen bat egin nahi izan dugu euskarazko fraseologia konputazionalaren arloan. Zehazki, euskarazko `izena+aditza` egiturako UFak corpusetik automatikoki erauzi eta idiomatikotasun-mailaren arabera karakterizatzeko lan esperimentala egin dugu.

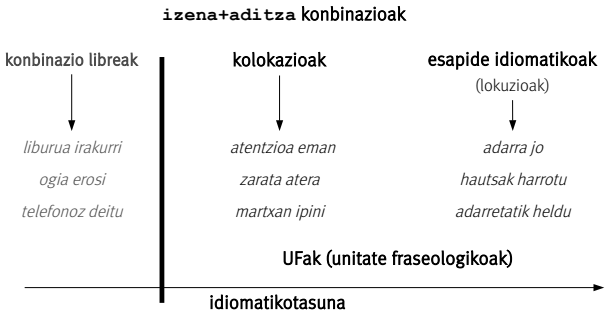
2 Arloaren egoera eta ikerketaren helburuak

2.1 Idiomatikotasunaren eta haren karakterizazioaren marko teorikoa

Idiomatikotasuna ez-konposizionaltasun semantikoarekin identifikatu izan da (Manning eta Schütze, 1999), baina, azken ikerkuntza batzuen arabera (Granger eta Paquot, 2008; Baldwin eta Kim, 2010), zenbait ezaugarri edo propietateren konbinazio konplexua da:

- Instituzionalizazioa (idiosinkrasia estatistikoa). Hiztunek unitatetzat hautematen dute konbinazioa. Osagaiek halako joera bat dute elkarrekin konbinatzeko, eta emaitza "ohikoa" edo "ezaguna" da.
- 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').
- Finkapen morfosintaktikoa: konbinazioa ezin da egitura morfosintaktiko guztietan erabili. *Adarra jo* esapidean, *adar* izena ez da pluralean erabiltzen, eta *adarrak ipini* unitatean, berriz, ezin da singularra erabili.
- Finkapen lexikala: osagai bat edo batzuk ezin dira sinonimoez ordezkatu. Esaterako: *ziria sartu*, baina *zotza sartu*?, *ezkarda sartu*?; gastuak murrizteaz mintzatuz, *gerrikoa estutu* diogu, baina *uhala estutu*?

Aurreko ezaugarr irizpideak ezartzearen emaitza continuum b ıum horretan "eremu" edo zona desberdinal



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 (Heid, 2008). Joera nagusia da horiek erdikonposizionalak, lexikalki murriztuak eta morfosintaktikoki aski malguak direla esatea.

2.2 UFen karakterizazio automatikoa

Karakterizazioaren aurreko ezinbesteko urratsa hautagaiak erauztea da. Horretarako sistema batek prozesamendu linguistiko minimo bat izan behar du, eta, oro har, esan daiteke lematizazioa eta etiketatze morfosintaktikoa direla minimo hori adierazten duten estandarrak (Heid, 2008).

UF hautagaiak karakterizatzeko, zenbait teknika proposatu dira literaturan. UFen lau ezaugarrietako bakoitza neurtzeko, teknika hauek erabili ohi dira, hurrenez hurren:

- Agerkidetza: konbinazioaren estatistikak osagaiak ausaz konbinatuko balira espero litezkeen estatistikekin konparatzea, elkartze-neurrien bidez.
- Antzekotasun distribuzionala: konbinazioaren testuinguruak osagaien testuinguruekin konparatzea.
- Konbinazioaren portaera morfosintaktikoaren eta egitura bereko konbinazioen batez besteko portaeraren arteko distantzia.

- Osagaien ordezkagarritasuna: konbinazioaren estatistikak osagaiak sinonimoez ordezkatuz sortzen diren konbinazioen estatistikekin konparatzea.

Konparazio horien emaitza zenbat eta desberdinago, hautagaia hainbat eta idiomatikoago. Agerkidetza da gehien erabili den teknika, eta hori neurtzeko elkartze-neurriak hasieratik erabili dira arlo honetako ikerkuntzan (Church eta Hanks, 1990; Smadja, 1993). Beste ezaugarriak kuantifikatzeko ikerlanak nabarmen ugaritu dira azken hamarkadan (Baldwin *et al.*, 2003; Van de Cruys eta Moirón, 2007; Fazly eta Stevenson, 2007), dela hautagaien rankingak osatzeko, dela automatikoki sailkatzeko.

2.3 Ikerketaren helburuak

Ikerkuntza-lan honen helburu nagusia izan da corpusetatik `izena+aditza` osaerako UFak automatikoki eskuratzeko eta haien idiomatikotasunaren arabera karakterizatzeko teknikak garatzea, eta ikergai hauen inguruko ezagutza eskuratzea eta sortzea: a) idiomatikotasunaren eta haren propietate bakoitzaren neurketen artean dagoen korrelazioa; b) UFen propietateen ebidentzia enpirikoak zenbateraino datozen bat teoria fraseologikoen aurresanekin; eta c) idiomatikotasuna konplexua izanik, propietateen neurketak konbinatuz emaitza hobeak lortzen diren ala ez jakitea.

3 Lan esperimentala

3.1 Lan esperimentalaren diseinua

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

1	$\texttt{izena}_{\text{subjektua}}$ + `aditza`: *burua joan, eguzkia sartu*; *gogoak eman, loak hartu*
2	$\texttt{izena}_{\text{objektua}}$ + `aditza`: *hanka sartu, zubiak eraiki, lan egin, itxurak egin, zarata atera, gola sartu*
3	$\texttt{izena}_{\text{subjektuaren pred.}}$ + `aditza`: *beldur izan, falta izan, giro egon*
4	$\texttt{izena}_{\text{objektuaren pred.}}$ + `aditza`: *atsegin ukan/*edun, damu ukan/*edun*
5	$\texttt{izena}_{\text{datiboa}}$ + `aditza`: *edanari eman, bideari ekin, lanari lotu*
6	$\texttt{izena}_{\text{adjuntua}}$ + `aditza`: *mendean hartu, adarretatik heldu, harira etorri, aurrez ikusi*

1 Taula: `izena+aditza` osaerako konbinazio-moten errepertorioa.

Bi karakterizazio-ataza bereizi dira: a) ranking-ataza, non propietate bakoitzaren neurketen emaitzak hautagaiak ordenatzeko erabili baitira; eta b) sailkatze-ataza, non esperimentu bakunen emaitzak konbinatu baititugu, ikasketa automatikoko teknikak erabiliz. Esperimentuetarako, 74 milioi hitzeko kazetaritza-corpus bat erabili dugu, Euskaldunon Egunkariako eta Berriako artikuluez osatua.

3.2 UF hautagaiak erauztea

UF hautagaiak testutik automatikoki erauzteko eta forma kanonikoa esleitzeko prozesu automatikoa garatu da. Erauzketa-prozesuaren deskribapen zehatza Gurrutxaga eta Alegria (2011) lanean egin da. Bi urratsetan antolatua dago: bigrama-sorkuntza eta bigramen forma kanonikoa lortzeko normalizazioa. Lehen urratserako, Ngram Statistics Package erabili dugu[1]. Erauzketaren aurretik, corpusa linguistikoki prozesatu da, Ixa taldearen Eustagger etiketatzailearen bidez lehenik (Alegria *et al.*, 2002), eta, ondoren, ikerkuntza honetan landutako prozesu batzuk inplementatu dira, corpusak erauzketarako behar diren ezaugarri batzuk izan ditzan.

[1] http://search.cpan.org/dist/Text-NSP

3.3 Ebaluazioa: metodologia eta baliabideak

Ranking bidezko atazan, heinen korrelazio-koefizienteak erabili ohi dira, eta guk Kendall τ aukeratu dugu, berdinketak kudeatzeko modalitatean (Kendall τ_B). Batez besteko doitasuna (*AP - Average Precision*) ere erabili dugu, kategoria bakoitzeko hautagaiekiko ebaluazioa egiteko ($AP_{\texttt{UF}}$, $AP_{\texttt{id}}$ eta $AP_{\texttt{col}}$). Sailkapen automatikoaren bidezko atazen ebaluazioan, Weka tresnak eskaintzen dituen zenbait neurri erabili ditugu: ondo sailkatutako instantzia-kopurua (CC), klase bakoitzaren F neurriak, F_{mikro} eta F_{makro}.

Ebaluaziorako *gold standard*tzat, erauzketatik 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 arteko adostasunerako, adostasun ertainaren mailako 0,55eko Fleiss κ lortu dugu (Landis eta Koch, 1977). Horrez gain, erauzketaren parametro batzuk doitzeko, beste erreferentzia bat ere eratu dugu, iturri hauetako `izena+aditza` unitateez osatua: a) Euskaltzaindiaren *Hiztegi Batua*; b) Ibon Sarasolaren *Euskal Hiztegia*; c) Elhuyar Fundazioaren *Euskara–Castellano/Castellano–Vasco Hiztegia*; d) *Intza proiektua*[2]; eta e) Ixa taldearen EDBL datu-base lexikala (Aldezabal *et al.*, 2001).

3.4 Idiomatikotasunaren ezaugarrien edo propietateen banakako neurketa

3.4.1 Idiosinkrasia estatistikoa, agerkidetza neurtuz

Bigramen agerkidetza-datuen analisi estatistikoa S. Everten UCS toolkit[3] paketearen bidez egin dugu (Evert, 2005). 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.

3.4.2 Konposizionaltasun semantikoa, antzekotasun distribuzionala neurtuz

Hauek dira UFen konposizionaltasuna karakterizatzeko gure metodologiaren oinarriak:

- UF hautagaien testuinguruak haren osagai bakunen testuinguruekin konparatzea.
- 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.

Antzekotasun distribuzionaleko tekniken bidez egin dugu testuinguruen konparazioa (Gurrutxaga eta Alegria, 2012). Lehenik, bektore-espazioaren ereduan (VSM – *Vector Space Model*) ohikoak diren neurri batzuk esperimentatu ditugu [Jaccard koefizientea, kosinua eta Jensen-Shannon dibergentzia; Berry-Roggheren (1974) R balioa eta Wulffek (2008) egindako bi hedapenak]. VSMren inplementazio berezia den ezkutuko semantikaren analisia (LSA – *Latent Semantic Analysis*) izeneko teknikaren aplikazioa ere esploratu dugu. Infomap softwarea erabili dugu horretarako[4]. Azkenik, dokumentu arteko antza kalkulatzeko IR arloko indize batzuk ere aplikatu ditugu. Lemur Toolkit aukeratu dugu testuinguru-dokumentuen arteko antza kalkulatzeko[5].

3.4.3 Malgutasun morfosintaktikoa, erreferentzia-portaerarekiko distantzia neurtuz

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

- Portaera orokorrarekiko konparazioa (Barkema, 1994; Fazly eta Stevenson, 2007; Wulff, 2008): 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.

[2] http://intza.armiarma.com
[3] http://www.collocations.de/software.html
[4] http://infomap-nlp.sourceforge.net/
[5] http://www.lemurproject.org

Kontuan hartutako aldakuntzak. Malgutasun morfosintaktikoa neurtzeko aztertzen diren aldakuntzen multzoa hizkuntzaren ezaugarrien araberakoa da. Arlo honetako ikerlan esanguratsu batzuk (Martínez, 1996; Oyharçabal, 2006; Odriozola, 2010) eta EDBLko aditz-lokuzioen gauzatze-eskemak (Urizar, 2012) kontuan izanik, kasu hauek hautatu ditugu malgutasun morfosintaktikoa karakterizatzeko:

- Izenaren ezker- eta eskuin-hedapenak: determinatzailea (*liburu* ***bat*** *irakurri dut*); izenondoa (*liburu* ***interesgarria*** *irakurri nuen*); izenlaguna (***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.*
- 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.*
- Osagaien ordena (`IZE ADI` / `ADI IZE`): *liburua irakurri dut/irakurri dut liburua.*

Bigramaren portaera ezagutzeko, aldakuntza bakoitzean duen maiztasuna kontatzen dugu. Bestetik, malgutasuna kalkulatzeko erabili ditugun bi konparazio-prozeduretako portaera ezagutzeko behar diren kontaketak ere egiten dira. Neurri hauek erabili ditugu portaeren arteko distantzia kalkulatzeko: Kullback-Leibler dibergentzia (Fazly eta Stevenson, 2007); Wulffen (2008) SSD neurria (*sum of squared deviations*) eta entropia erlatiboa; eta Bannarden (2007) CPMI (*conditional pointwise mutual information*).

3.4.4 Malgutasun lexikala, osagaien ordezkagarritasuna neurtuz

Bigramen osagaien ordezkagarritasuna neurtzeko, osagaien ordezkoak behar ditugu. Horretarako baliabide hauek erabili ditugu: Elhuyarren *Sinonimoen Kutxa*[6]; eta Ixa taldearen Euskal WordNet (Pociello *et al.*, 2011). Halaber, baliabideon estaldura handitzeko, konbinatu egin ditugu, baita hedatu ere, bigramaren osagai bakoitzaren Euskal WordNeteko ahaideak (*siblings*) gehituz.

Ordezkagarritasuna neurtzeko, Van de Cruys eta Moirónen (2007) R_{nv} eta R_{vn} indizeak erabili ditugu, eta Fazly eta Stevensonen (2007) $\text{Fixedness}_{\text{lex}}$ neurria.

3.4.5 Esperimentu bakunen emaitzen analisia

2 taulan laburbildu ditugu teknika bakoitzarekin lortutako emaitza onenak. 2 irudian, horien P/R (doitasuna/estaldura) kurbak.

2 Taula: Esperimentu bakunetan teknika bakoitzarekin lortutako emaitza onenak.

teknika	τ_B	$AP_{\texttt{UF}}$	$AP_{\texttt{id}}$	$AP_{\texttt{col}}$
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

Emaitza horien alderdi esanguratsuenak:

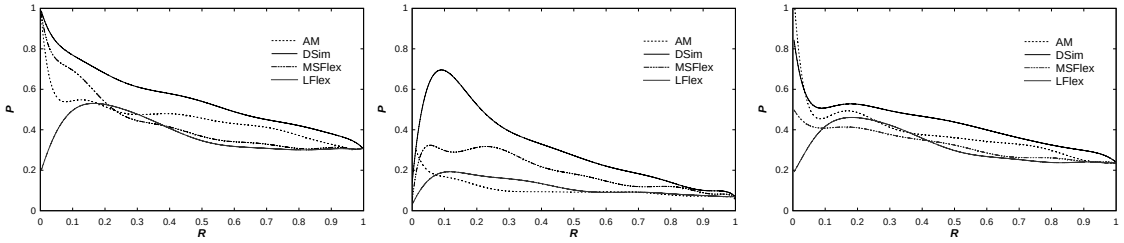
- Antzekotasun distribuzionaleko neurriak (DSim) dira idiomatikotasun-mailarekin korrelazio onena dutenak[7]. Bestetik, esapide idiomatikoen erauzketan da are nabariagoa horien nagusitasuna.
- Kolokazio-erauzketan, AM neurri onena (t neurria) DSim neurrien lehiakide handia da, baina $AP_{\texttt{col}}$ emaitza onena lortu duena aditzaren semantika neurtzen duen Indri indize bat izan da. Emaitza hori oso garrantzitsua da, kolokazioak erdikonposizionalak direlako ikuspegiarekin bat baitator.

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

[7] Zehazki, testuinguru-dokumentuen arteko antza neurtzen duten Indri indizea eta KL dibergentzia dira neurri onenak, L2 esperimentu-modalitatean.

- MSFlex emaitza onenak esapide idiomatikoekin lortu dira ($AP_{\texttt{id}}$; zehazki, mugatasun-aldakuntzen neurketetan). Gainerakoan, ez dituzte agerkidetza-esperimentuen emaitzak gainditzen.
- LFlex neurketetan, emaitzak espero baino txarragoak dira.

2 Irudia: Idiomatikotasun-rankingen P/R (doitasuna/estaldura) kurbak.



(a) UFak (b) Esapide idiomatikoak (c) Kolokazioak

3.5 Propietateen integrazioa: ikasketa automatikoa

Ikasketa automatikoko esperimentuak egiteko, Weka paketea erabili dugu[8]. Sei algoritmorekin probak egin ondoren (Gurrutxaga eta Alegria, 2013), bi hauekin lortu ditugu emaitza onenak[9]: Logistic Regression eta SMO (*Sequential Minimal Optimization*)[10].

Atributu gisa, aurreko ataleko neurketen emaitzak erabili ditugu, eta, Fazly eta Stevensoni (2007) jarraituz, konbinazioaren aditza ere bai. Ebaluazioan, balidazio gurutzatua erabili dugu, ebaluazio-erreferentzia ez delako oso handia (1 145 instantzia). Atributuak hautatzeko iragazkiak ikaste-multzoan dauden instantziak soilik azter ditzan, AttributeSelectedClassifier metasailkatzailea hautatu dugu.

3 taulan, lau datu-multzorekin lortutako emaitzak bistaratu ditugu: `DSim` (antzekotasun distribuzionaleko neurketak); `4 osag.` (lau propietateen neurketak); `4 osag.+ad.` (aurreko atributuak eta aditza); eta `CS-BF` (AttributeSelectedClassifier metasailkatzailean, CfsSubsetEval1 ebaluatzaileaz, eta BestFirst bilaketa-metodoaz osatutako iragazkia).

3 Taula: Ikasketa automatikoko esperimentuen emaitzak.

Atrib.	Metod.	CC	$F_{\texttt{id}}$	$F_{\texttt{col}}$	$F_{\texttt{free}}$	F_{mikro}	F_{makro}
Oinarri-lerroa		69,607	0,000	0,000	0,821	0,571	0,274
`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

Emaitza onenak SMO algoritmoarekin lortu dira, `4 osag.+ad.` datu-multzoa erabiliz. Beraz, ezaugarri guztiek egiten dute beren ekarpena. Hala ere, datu-multzo hori osatzeko prozesamendua handia da, eta LR metodoak atributu-hautaketa automatikoarekin lortutako emaitzek atea irekitzen diote bideragarritasunaren eta emaitzen kalitatearen arteko erlazio on bat lortzeko aukerari.

4 Ondorioak

Ondorio nagusia da ataza honetan estandar diren agerkidetza-tekniken emaitzak modu esanguratsuan gainditu direla, batez ere teknika semantikoen bidez, baina baita malgutasun morfosintaktikoaren neur-

[8] http://www.cs.waikato.ac.nz/ml/weka/
[9] Gainerako lauak: Naive Bayes, C4.5 decision tree (j48), RandomForest eta PART.
[10] *Support Vector Machine* edo sostengu-bektoreen makinen modalitatearen inplementazio bat.

ketaren bidez ere. Aldiz, finkapen lexikalaren neurketak ez du espero zen mailako emaitza izan.

Bestetik, fraseologiaren aurresan batzuen ebidentzia esperimentalak lortu ditugu: idiomatikotasunaren konplexutasuna, konposizionaltasunaren UF-kategorien zeharreko graduazioa, eta kolokazioen erdikonposizionaltasuna eta malgutasun morfosintaktikoa.

Ikerketa honen ekarpenak baliagarriak dira etorkizuneko hiztegigintzak automatizaziorantz izango duen bilakabidean, eta hizkuntzaren prozesamenduko arloko zenbait atazatan, hala nola datu-base lexikalen elikatzean, corpusen etiketatzean eta, testuinguru eleaniztunean aplikatuta, itzulpen automatikoan.

5 Etorkizuneko lanak

Lehenik, informazio sintaktiko aberatsagoa (adib., osagaien mendekotasunak) erabiltzeak erauzketa hobetu dezake. Bestetik, interesgarria da interpretazio literala eta idiomatikoa izan ditzaketen konbinazioen agerpenak bereiztea. Azkenik, komeni da garatutako metodologia corpus paraleloei aplikatzea, bi hizkuntzatako UF bikote baliokideak erauzteko eta karakterizatzeko.

Erreferentziak

Aldezabal, I., O. Ansa, B. Arrieta, X. Artola, A. Ezeiza, G. Hernández, eta M. Lersundi. 2001. EDBL: A general lexical basis for the automatic processing of Basque. In *IRCS Workshop on linguistic databases*, 1–10, Philadelphia, Pennsylvania.

Alegria, I., Maxux Aranzabe, Aitzol Ezeiza, Nerea Ezeiza, eta Ruben Urizar. 2002. Robustness and customisation in an analyser/lemmatiser for Basque. In *Proceedings of Workshop on "Customizing knowledge in NLP applications". Third International Conference on Language Resources and Evaluation*, 1–6, Las Palmas de Gran Canaria.

——, A. Gurrutxaga, P. Lizaso, X. Saralegi, S. Ugartetxea, eta R. Urizar. 2004. An xml-based term extraction tool for Basque. In *LREC2004: 4th International Conference On Language Resources And Evaluation*, 1733–1736, Lisboa.

——, A. Gurrutxaga, X. Saralegi, eta S. Ugartetxea. 2006. Elexbi, a basic tool for bilingual term extraction from Spanish-Basque parallel corpora. In *12th EURALEX International Congress*, 159–165, Turin.

Baldwin, T., C. Bannard, T. Tanaka, eta D. Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, volume 18, 89–96, Sapporo.

——, eta S.N. Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing, second edition*, ed. by N. Indurkhya eta F. J. Damerau, 267–292. Boca Raton, AEB: CRC Press, Taylor and Francis Group.

Bannard, C. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, 1–8, Praga. ACL.

Barkema, Henk. 1994. Determining the syntactic flexibility of idioms. *Realising and Using English Language Corpora* 39–52.

Berry-Rogghe, G.L.M. 1974. Automatic identification of phrasal verbs. *Computers in the Humanities* 16–26.

Church, K.W., eta P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16. 22–29.

Cowie, A.P. 1998. *Phraseology: Theory, Analysis, and Applications*. Oxford University Press, USA.

Evert, S., 2005. *The statistics of word cooccurrences: Word pairs and collocations*. University of Stuttgart tesia.

Fazly, A., eta S. Stevenson. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 9–16, Praga. ACL.

Fillmore, Charles J. 1979. On fluency. In *Individual Differences in Language Ability and Language Behavior*, ed. by D. Kempler eta W. S. Y. Wang, 85–101. New York: Academic Press.

Granger, S., eta M. Paquot. 2008. Disentangling the phraseological web. In *Phraseology: An Interdisciplinary Perspective*, ed. by S. Granger eta F. Meunier, 27–50. John Benjamins Publishing.

Gurrutxaga, A., eta I. Alegria. 2011. Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World. ACL HLT 2011*, 2–7, Portland, Oregon.

——, eta ——. 2012. Measuring the compositionality of NV expressions in Basque by means of distributional similarity techniques. In *Proceedings of the 8th international Conference on Language Resources and Evaluation - LREC 2012*, 2389–2394, Istanbul.

——, eta ——. 2013. Combining different features of idiomaticity for the automatic classification of noun+ verb expressions in Basque. In *Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013) NAACL HLT 2013*, volume 13, 116–125, Atlanta, Georgia.

Heid, U. 2008. Computational Phraseology. An overview. *Phraseology: an interdisciplinary perspective* 337–360.

Landis, J Richard, eta Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 159–174.

Manning, C. D., eta H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

Martínez, A, 1996. Syntactic evidence in favour of degrees of incorporation in [*n*+*egin*] constructions. Eskuizkribua.

Odriozola, J.C., 2010. Euskararen aditz-unitate fraseologikoen deskribapena. Unibertsitateko katedra plazarako lehiaketa. Zientzia eta Teknologia fakultatea.

Oyharçabal, Bernard. 2006. Basque light verb constructions. *Anuario del Seminario de Filología Vasca Julio de Urquijo. R. L. Trasken oroitzapenetan ikerketak euskalaritzaz eta hizkuntzalaritza historikoaz* 40. 787–806.

Pociello, E., E. Agirre, eta I. Aldezabal. 2011. Methodology and construction of the basque wordnet. In *Language resources and evaluation*, volume 45, 121–142. Springer.

Sag, I., T. Baldwin, F. Bond, A. Copestake, eta D. Flickinger. 2010. Multiword expressions: A pain in the neck for NLP. *Computational Linguistics and Intelligent Text Processing* 189–206.

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

Sinclair, J. 1996. The search for units of meaning. *Textus* 9. 75–106.

Smadja, F. 1993. Retrieving collocations from text: Xtract. *Computational linguistics* 19. 143–177.

Urizar, Ruben, 2012. *Euskal lokuzioen tratamendu konputazionala*. Donostia: Informatika Fakultatea, UPV/EHU tesia.

Van de Cruys, T., eta B.V. Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 25–32, Praga. ACL.

Wulff, S. 2008. *Rethinking Idiomaticity*. Corpus and Discourse. New York: Continuum International Publishing Group Ltd.

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). Ikerkuntzaren emaitzak MWE 2011 (ACL), LREC 2012 eta MWE 2013 (NAACL) nazioarteko biltzarretan aurkeztu dira (Gurrutxaga eta Alegria, 2011, 2012, 2013).

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