Unitate fraseologikoen agerpen literalak, *urre baina urri*

Uxoa Iñurrieta

IXA taldea, HiTZ. Euskal Herriko Unibertsitatea.

usoa.inurrieta@ehu.eus

Laburpena

Unitate fraseologiko asko idiomatikoki eta literalki uler daitezke. Esate baterako, *ziria sartzea*k bi esanahi izan ditzake testuinguruaren arabera: norbaiti iruzur egitea edo nonbait ziri bat sartzea literalki. Lan honetan, corpusetan oinarritutako azterketa eleaniztun baten berri emango dugu, eta erakutsiko dugu, batetik, halako hitz-konbinazioak oso gutxitan erabiltzen direla literalki praktikan, eta bestetik, idiomatiko-literal bereizketa egiteko garrantzitsua dela semantika ez ezik morfosintaxia ere kontuan hartzea. Lan hau azterketa zabalago batetik (Savary *et al.*, 2019) eratorria da; bost familiatako hizkuntza bana hartu da kontuan azterketa egiteko, eta horietako bati eskainiko diogu arreta berezia hemen: euskarari.

Hitz gakoak: fraseologia, idiomatikotasuna, literaltasuna, hizkuntzalaritza konputazionala

Abstract

Multiword expressions can be understood either idiomatically or literally. For instance, the Basque expression 'ziria sartu' (lit. 'to put a spike in') can have two meanings: to trick someone or to literally put a spike somewhere. This paper presents a corpus-based analysis carried out in five typologically different languages, with special focus on the Basque language. We show, on the one hand, that literal occurrences of these kinds of word combinations constitute a rare phenomenon, and on the other hand, that not only semantics but also morphosyntax is helpful to distinguish one meaning from the other. This is a derivative piece of work of (Savary et al.*, 2019).*

Keywords: *phraseology, idiomaticity, literalism, computational linguistics*

1 Sarrera

Unitate Fraseologiko (UF) deritze nolabait idiomatikoak diren hitz-konbinazioei, hau da, hitz batez baino gehiagoz osaturik egon arren erabat konposizionalak ez direnei (Baldwin eta Kim, 2010). Halako hitz-konbinazioak idiomatikoki nahiz literalki uler daitezke askotan, eta testuinguruaren arabera argitu ohi da esanahi batez ala besteaz ari garen. Esate baterako, *ziri* eta *sartu* hitzak idiomatikoki erabilita daude 1. adibidean, baina literalki 2.ean[1].

(1) *Ez zen benetan ari.* ***Ziria sartu*** *zizun!*

(2) *Mutikoak egurrezko ziria sartu zuen zuloan.*

Esanahi idiomatikoen eta literalen arteko bereizketak hizkuntzalaritzako eta psikolinguistikako hainbat ikertzaileren arreta piztu du, eta Hizkuntzaren Prozesamenduko (HP) erronkarik handienetakotzat hartzen da gaur egun, halakoak konputazionalki desberdintzea ertz askoko lana baita (Constant *et al.*, 2017). Hain zuzen ere, PARSEME proiektu europarrak (Savary *et al.*, 2015) fraseologia konputazionalaren arloko ikertzaileak bildu nahi izan ditu, UFek HPn sortzen dituzten zailtasunei nola aurre egin ikertzeko. Proiektu horren baitan, hogei hizkuntzatako corpusak etiketatu dira fraseologia mailan, tartean euskarazkoa, eta corpus horixe erabili dugu guk artikulu honen oinarri den azterketa burutzeko. Oraingoan, ordea, UFei beharrean, UFen agerpen literalei begiratu diegu. Bi hipotesi nagusi hartu ditugu abiapuntutzat: lehena, UFak teorian idiomatikoki nahiz literalki erabil badaitezke ere, praktikan oso gutxitan erabiltzen direla literalki; eta bigarrena, UFak eta beren agerpen literalak bereizteko garrantzitsua dela semantikari ez ezik morfosintaxiari ere begiratzea.

[1] Adibideetan, letra lodiz markatuko ditugu agerpen idiomatikoak, eta azpimarratuta agerpen literalak. Kointzidentziazko agerpenei (ikus azalpena 4.2. atalean) eta gure aztergaitik kanpoko adibideei, berriz, traola jarriko diegu aurretik.

Izan ere, UF asko malguak izaten dira morfosintaxiari dagokionez, baina beste askok murriztapenak izaten dituzte (*min eman* edo *mina eman*, baina *aurre egin* eta ez **aurrea egin*). Murriztapen horiek, hain zuzen, lagungarriak dira UF asko agerpen literaletatik bereizteko. Lehen eta bigarren adibideetan, nahikoa da jakitea *ziria sartu* UFak ez duela modifikatzailerik izaten; bigarren esaldiko izen-sintagman *egurrezko* adjektiboa ageri direnez, erraz bereiz liteke agerpen hori ez dela idiomatikoa.

Tipologia desberdineko bost hizkuntza landu ditugu agerpen literalen inguruko azterketa zabal baten bidez (Savary *et al.*, 2019): alemana, greziera, euskara, poloniera eta portugesa. Azterketa horren ideia nagusiak ekarriko ditugu hona, baina garrantzia euskarazko zatiari emanez. Hasteko, UFen agerpen literalen inguruan zer ikerketa egin den azalduko dugu (2. atala). Gero, PARSEMEren corpusaz eta hura etiketatzeko irizpide nagusiez jardungo dugu (3. atala), eta agerpen literalen inguruko azalpenei ekingo diegu jarraian (4. atala). Zehazki zer eta nola aztertu dugun argitu ondoren, lortutako emaitzak erakutsiko ditugu, eta gako batzuk emango ditugu emaitza horiek interpretatzeko (5. atala). Azkenik, ondorio nagusiak ezagutarazi, eta etorkizuneko lan-ildoak aurkeztuko ditugu (6. atala).

2 Arloa zertan den

UFen agerpen idiomatikoen eta literalen arteko bereizketa hainbat ikuspuntutatik landu izan da, psikolinguistikatik hasi eta hizkuntzalaritza konputazionalera, hizkuntzalaritzako azterketa deskriptiboak ere tarteko direla. Psikolinguistikan, esaterako, UFak giza garunean nola gordetzen eta prozesatzen diren ikertu izan da (Cacciari eta Corradini, 2015), agerpen idiomatikoak eta literalak parez pare jarrita. Ikertzaile batzuek, Geeraert *et al.*-ek adibidez (2018), begien mugimenduak neurtuz interpretatu dute hiztunok zer sentipen dugun bi agerpen mota horiekiko, eta atera duten ondorio nagusia da UFen agerpen literalak arrotzak egiten zaizkigula hiztunoi oro har. Geroago erakutsiko dugunez, ondorio hori bat dator gure lanarekin, gure azterketaren arabera agerpen literalak oso gutxitan agertzen baitira corpus errealetan.

Bestetik, UFen tipologiak proposatzean ere sakon aztertu izan dira agerpen idiomatikoen eta literalen arteko aldeak. Sheinfux *et al.*-en lanean (2017), esaten da UF mota batzuek beste batzuek baino joera handiagoa dutela agerpen literalak izateko eta, gainera, UF moten malgutasun morfosintaktikoa ere aurresangarria dela askotan. Haien esanetan, esapide figuratiboen (adib.: *arrastoa utzi*) agerpen literalak natural samarrak egiten zaizkigu, baina ez esapide erabat opakoenak (adib.: *adarra jo*). Era berean, egileek diotenez, esapide figuratiboak malguagoak dira morfosintaxi aldetik ez-figuratiboak baino (*arrasto luzea utzi*, baina ez *adar ederra jo*).

Gramatiken arloan ere landu da gaia, gramatika formaletan bereziki (Abeillé eta Schabes, 1989). Euskarari dagokionez, gramatikek eta lan deskriptiboek nahiko zabal aztertu dituzte oso ohikoak diren UF batzuk, beste lan batzuetan *aditz-esapide* edo *aditz konplexu* ere deitu izan zaienak hain zuzen (Hualde eta Ortiz de Urbina, 2003, 223–227, 235–246, 307–308. orr.): *ari izan*, *behar izan*, *min hartu/eman* eta gisa horretakoak. Halakoak askoz ere ugariagoak dira gurean inguruko hizkuntzetan baino (Inurrieta *et al.*, 2018), eta ziur asko horregatik eskaintzen zaie arreta berezia euskal literatura fraseologikoan. Zabalak (2004), adibidez, aditz arinak barne hartzen dituzten UFen deskribapen xehea egiten du bere lanean, ezaugarri semantikoak eta morfosintaktikoak kontuan harturik. Dena dela, lan horietan ez da askorik esaten UFen agerpen literalez; UFen osaerari eta testuan izaten duten portaerari eman ohi zaie garrantzia.

Hizkuntzaren Prozesamendura etorrita, lan asko egin da UFen inguruan, erronka zaila baita halakoak konputazionalki ondo tratatzea (Constant *et al.*, 2017). Agerpen idiomatikoen eta literalen arteko bereizketak sekulako garrantzia du HPn, bereizketa horren araberakoa baita tresna automatiko askoren kalitatea. Euskaraz ere egin da lanik fraseologia konputazionalaren arloan, dela corpusetatik UFak erauzi eta hiztegietan sartzeari begira (Gurrutxaga *et al.*, 2014), dela analisi sintaktikoetan halako hitz-konbinazioak nola tratatu erabakitzeko (Urizar, 2012). Bigarren ataza horretan bereziki, UFak beren agerpen literaletatik bereiztean datza gakoa, bai eta itzulpengintza automatikoan ere, hala erabakitzen baita hitz-konbinazio jakin bati ordain idiomatikoa ala literala eman (Inurrieta *et al.*, 2017).

Azterketa kuantitatiboei dagokienez, aldiz, ezer gutxi egin da orain arte, guk dakigula. Halako bi lan egin dira poloniera oinarritzat hartuta: batek dio idiomatikotasun-tasa % 95ekoa dela aditzak, izenak, adjektiboak eta adberbioak burutzat dituzten UFetan (Waszczuk *et al.*, 2016); besteak, berriz, aditz-UFak bakarrik hartzen ditu aztergaitzat (Savary eta Cordeiro, 2017), eta agerpenen % 98 idiomatikoak direla ondorioztatzen du. Bigarren horri jarraituz egin dugu guk artikulu honetan azalduko dugun lana, baina hizkuntza gehiagotara zabaldu dugu azterketa, erakusteko ondorio horiek polonieratik haragokoak direla.

3 UFen inguruko kontzeptu nagusiak eta PARSEMEren corpusa

Sarreran esan dugunez, PARSEME proiektu europarrean sortutako corpus elaniztuna hartu dugu oinarritzat UFen agerpen literalak aztertzeko. Corpus horretan zer-nolako etiketak dauden azaltzeko, labur ditzagun orain etiketatze-gidalerroetako kontzeptu nagusiak (3.1. azpiatala), corpusean bereizten diren UF moten ezaugarriak (3.2. azpiatala) eta corpus etiketatuari buruzko datu orokorrak (3.3. atala). Xehetasun gehiago behar dituenak Savary *et al.*-en lanean (2018) ditu eskuragarri gidalerroak osorik[2], eta Inurrieta *et al.*-enean (2018) euskarazko corpusari buruzko argibideak eta gogoetak.

3.1 Etiketatze-irizpideak

PARSEMEren corpusean **aditz-UFak** daude etiketatuta, hau da, buru sintaktikotzat aditza duten hitz-konbinazio idiomatikoak. Horrek esan nahi du, 3. adibidea etiketatzekoa bada ere, 4.eko izen elkartua ez dela kontuan hartzen[3], izena duelako buru sintaktikotzat eta ez aditza.

(3) ***Izena eman*** *zuen ikastaroan.*

(4) # *Izen-ematea atzo amaitu zen.*

Aditz nagusiarekin batera agertzen diren hitzek edozein kategoria gramatikal izan dezakete. Esate baterako, 3. adibideko izena+aditza konbinazioaz gainera, etiketatuta daude 5. adibideko adjektiboa+aditza konbinazioa eta 6.eko perpaus koordinatua ere.

(5) ***Nabari da*** *jendea etorri dela.*

(6) ***Ikusi eta ikasi***!

UFen barruko **osagai lexikalizatuak** dira markatzen direnak, hau da, UFan beti agertzen diren lemei dagozkien hitz-formak. Nolanahi ere, euskararen izaera aglutinatiboa dela-eta, lexikalizatu gabeko morfema batzuk ere etiketatuta daude euskaraz, lexikalizatutako lemaren bati lotuta zeudelako. Adibidez, 7–9. adibideetan *pauso* eta *eman* dira UFko osagai lexikalizatuak, baina, etiketatzea hitz mailan egina denez, izenari itsatsitako markak ere etiketaren barruan sartu behar izan dira nahitaez: 7. adibidean artikulua (*-ak*), eta 9.ean artikulua eta postposizio instrumentala (*-ez*).

(7) ***Pausoak ematen*** *ari da.*

(8) *Hainbat* ***pauso*** *oker* ***eman*** *zituen.*

(9) ***Emandako pausoez*** *damutu zen.*

UFen aldaera morfosintaktikoak kontuan hartzen dira PARSEMEren corpusean, eta horregatik dago etiketatuta *pauso eman* UFa adibide horietan guztietan, esaldi batetik bestera desberdin erabilita egon arren. Lehen bi adibideetan (7–8) aditza da buru sintaktikoa, eta azkenekoan, ordea, ez. Hala ere, kontuan hartu behar da UFen **forma kanonikoa**n pentsatu behar dela hitz-konbinazio jakin bat UFa den ala ez erabakitzeko, zenbaitetan posible baita UFetako osagaien erlazio sintaktikoa aldagarria izatea –erlatibozko perpausetan (9. adibidea) eta halakoetan–. Hau diote forma kanonikoei buruz PARSEMEren gidalerroetan: "Aditz-UF baten forma kanonikoa aditz-sintagma bat da, zeina boz aktiboan baitago, buru sintaktikotzat aditza baitu eta gainerako osagai lexikalizatuak aditzaren mende edo beste osagai lexikalizaturen baten mende baitaude."

Hori aintzat harturik, 9. adibideko konbinazioa forma kanonikora ekarrita, *pausoak eman zituen* litzateke, eta beteko luke burutzat aditza izateko baldintza. Ez da gauza bera gertatzen, ordea, 4.eko hitz elkartuarekin, adibide horretan hitz-konbinazioa forma kanonikoan baitago dagoeneko, izena buru duelarik: ez da *izen eman*, baizik eta *izen-emate*.

Beraz, 7. eta 9. adibideen bidez erakutsi dugunez, lexikalizatu gabeko morfema batzuk markatuta daude euskaraz. Hala ere, kontrakoa ere gertatzen da zenbaitetan: morfema lexikalizatu bat ez etiketatu izana, lexikalizatuta ez

[2]Sarean ere badago gidalerroen bertsio oso bat: `http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/?page=home`

[3]PARSEMEren gidalerroekin lotura zuzenik ez badu ere, merezi du aipatzea Azkarateren doktoretza-tesian (1987, 407–415. orr) badela eranskin oso bat hitz elkartuen eta aditz-UFen –haren hitzetan *aditz-esapideen*– arteko desberdintasunen inguruan.

dagoen lema bati itsatsita joateagatik. Esate baterako, 10. adibidean, *falta* eta *sumatu* bakarrik leudeke etiketatuta, *haren* gabe, nahiz eta *(norbaiten) falta sumatu* UFak derrigorrezkoa izan genitibodun osagai bat.

(10) *Haren **falta sumatzen** dut.*

Etiketak hitz mailan ematen direnez, ezin da *haren* ere markatu, horrek esan nahi bailuke *hura* lexikalizatutzat hartzen dela, eta ez da hala, lema hori ordezka baitaiteke beste askorekin: *norbaiten/zure falta sumatu*, eta abar.

Hori guztia gogoan izanik, ikus dezagun zer bi multzo bereizten diren euskaraz etiketatutako UFen artean.

3.2 Aditz-UFen sailkapena

PARSEMEren gidalerroak unibertsalak izateko asmoz sortu dira, hizkuntzaz hizkuntzako tradizio fraseologikoak nolabait uztartu eta proposamen bateratu bat sortze aldera. Hortaz, garrantzitsua da nabarmentzea gidalerroetako edukiak ez direla beti bateragarriak hizkuntzaz hizkuntzako literaturarekin. Guri dagokigunez zehazki, haien sailkapena ez dator guztiz bat euskaraz egin izan direnekin –ikus Gurrutxagak (2014) eta Urizarrek (2012, 78–71. orr) eginiko sailkapenak–, bi alderditan bereziki: aditz arindunak (ikus azalpena beherago) ez diren kolokazioak (*deia jaso*, *ados egon* eta halakoak) ez direlako kontuan hartzen PARSEMEn, eta, beste lanen batean ere agerian jarri dugunez (Inurrieta *et al.*, 2018), aditz arindun konbinazioak izena+aditza konbinazioetara bakarrik mugatzen dituztelako, *korrika egin*, *nahiago izan* eta antzerakoak alde batera utzita.

Lan honetako edukiak PARSEMEren corpusarekin duenez zerikusia, haien sailkapenari lotuko gatzaizkio hemen. Sailkapen horrek sei aditz-UF mota bereizten ditu guztira, baina sei mota horietako bi baino ez dira hizkuntza guztietarako aplikagarriak. Bi horiek bakarrik dauzkagu, hain zuzen, euskaraz: aditz arindun konbinazioak (*Light Verb Constructions*, LVC) eta aditz-esapide idiomatikoak (*Verbal Idioms*, VID).

- **Aditz arindun konbinazioak** aditz batez eta izen batez osatuta egon ohi dira, eta izenak ematen dio hitzkonbinazioari esanahiaren zatirik handiena. Hau da: izenak ekintza, gertaera edo egoera bat adierazi ohi du, eta aditzak, normalean, ezaugarri morfologikoak bakarrik gehitzen dizkio: pertsona, numeroa, denbora edota aspektua. Izena aditzaren mendekoa izaten da beti, eta zenbaitetan artikulua, kasu-marka edo postposizioren bat izan dezake. Multzo horretakoak dira, adibidez, *lo egin, aurrera egin, min hartu* eta *negar egin*.
- **Aditz-esapide idiomatiko**ek, berriz, bi osagai lexikalizatu dituzte gutxienez, aditz bat eta haren mendeko osagai bat, baina mendeko osagai hori askotarikoa izan daiteke, eta ez, LVCetan bezala, izena bakarrik. Esanahiari dagokionez, konbinazioaren esanahi osoa ez da osagaien esanahien batura izaten: zenbaitetan ezin izaten da konbinazioa ulertu, aurretik jakin ezean zer esan nahi duen, eta beste batzuetan metafora bidez uler daiteke. Esate baterako, VID multzokoak dira *adarra jo, katuak mingaina jan* eta *begi-bistatik galdu*.

3.3 PARSEMEren corpusaren datu estatistikoak

Euskarazko corpus etiketatuak[4] bi iturritako testuak biltzen ditu: Dependentzia Unibertsalen corpuseko 6.621 esaldi, hau da, corpus osoa (Aranzabe *et al.*, 2019), eta Elhuyar Web Corpuseko[5] 4.537 esaldi. Hortaz, testu periodistikoetako eta sareko testuetako 11.158 esaldi ditu guztira, 157.807 hitz. Corpusari buruzko xehetasunak 1. taulan daude jasota, etiketa bakoitzeko kopuruak barne.

1. Taula: PARSEMEko euskarazko corpusaren datuak

Esaldiak	Hitzak	UFak	LVC	VID
11.158	157.807	3.823	3.049	774

Datozen ataletan azalduko dugun azterketak taula horretan jasotako etiketak ditu, hain zuzen, oinarritzat. Argitu dezagun orain nola erabili ditugun UF etiketa horiek agerpen literalak bilatzeko eta aztertzeko.

4 Zer aztertu dugun eta nola

PARSEMEren corpusa abiapuntutzat hartuta, corpusetik *hautagaiak* erauzi ditugu automatikoki, hau da, hainbat irizpideren arabera UFen agerpen literalak izan litezkeenak. Jarraian azalduko dugu nola erauzi ditugun hautagaiak

[4]Corpus osoa eskuragarri dago helbide honetan: `https://gitlab.com/parseme/sharedtask-data/tree/master/1.1`.
[5]`http://webcorpusak.elhuyar.eus/index.html`

zehazki (4.1. azpiatala), bai eta hautagai horiek nola multzokatu ditugun ere (4.2. azpiatala).

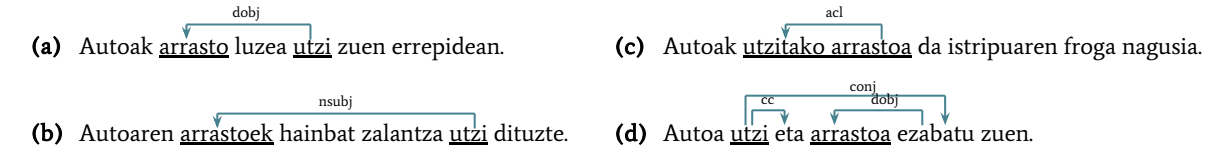
4.1 Hautagaiak lortzeko heuristikoak

Lan honen hasieran esan dugunez, ezaugarri morfosintaktikoek garrantzi handia dute, gure ustez, UFen agerpen idiomatikoak agerpen literaletatik bereizteko, eta, hortaz, morfosintaxi mailan etiketatutako corpus bat hartu dugu oinarritzat. Hala, corpusean bilaketak egitean, jakin ahal izan dugu lema bakoitzak zer kategoria gramatikal duen, zer beste morfema duen erantsirik (adibidez, kasu- edo postposizio-markak), eta esaldi bakoitzean zer erlazio sintaktiko duen beste lemekin. Aranzabe *et al.*-en lanean (2019) zehaztasun handiz dago azalduta nola egiten den analisi-prozesu hori.

PARSEMEren corpusa hutsetik aztertzen hasi beharrean, lau heuristikoren bidez erauzi ditugu agerpen literalak izan litezkeen hautagaiak, etiketatze-lana errazteko. Heuristiko horien atzean dagoen ideia zera da: UF baten barruko lemak corpusean elkarrekin agertu badira baina ez badira UFtzat etiketatu, baliteke hitz-konbinazio jakin hori jatorrizko UFaren agerpen literal bat izatea.

Demagun *arrastoa utzi* UFa etiketatuta dagoela corpusean, eta *arrasto* izena *utzi* aditzaren objektutzat agertzen dela etiketa horretan. Demagun, era berean, 1. irudiko lau esaldiak ere badaudela corpusean, baina *arrasto* eta *utzi* lemak ez daudela UFtzat etiketatuta. Heuristikoek esaldiz esaldi begiratuko lukete ea lema horiek baldintza jakin batzuk betetzen dituzten, eta hala erabakiko lukete hautagaiak erauzi ala ez. Beheko zerrendan jaso dugu hitz-konbinazio jakin batek zer baldintza bete behar dituen erauzia izateko, eta zer egingo lukeen heuristiko bakoitzak irudiko esaldiekin[6].

1. Irudia: Heuristikoek *arrastoa utzi* UFaren AL hautagaitzat erauzten dituzten lau adibide.



- **WindowGap**: lemek lau hitzeko leiho baten barruan agertu behar dute testuan, elkarren artean gehienez ere bi hitz dituztelarik.
  → Hautagai bana erauziko luke lau esaldietatik, lauretan agertzen baitira *arrasto* eta *utzi* lemak tartean gehienez ere bi hitz dituztela.
- **BagOfDeps**: lemek elkarri lotuta agertu behar dute dependentzia-zuhaitzean, baina berdin du zein hurrenkeratan dauden eta zer erlazio mota duten elkarren artean.
  → Hautagai bana erauziko luke (a), (b) eta (c) esaldietatik, baina ez (d) esaldidik, azken horretan *utzi* eta *arrasto* ez baitaude zuzenean elkarri lotuta dependentzia-zuhaitzean.
- **UnlabeledDeps**: lemek zuzenean elkarri lotuta agertu behar dute dependentzia-zuhaitzean, eta jatorrizko UFaren osagaien noranzko berean.
  → Lehen bi esaldietatik hautagaiak erauziko lituzke, baina ez beste bietatik. Izan ere, (a) eta (b) esaldietan bakarrik agertzen dira bi lemak elkarri lotuta eta aditza izenaren gobernatzaile delarik, hau da, dependentzia-erlazioari dagokion gezia aditzetik izenera doalarik.
- **LabeledDeps**: lemek elkarri lotuta agertu behar dute dependentzia-zuhaitzean, jatorrizko UFaren osagaien noranzko berean eta erlazio mota berberarekin.
  → Lau esaldietatik hautagai bakarra erauziko luke, (a) esaldian bakarrik agertzen baita *arrasto* lema *utzi*ren objektu zuzentzat.

Pentsatzekoa denez, zerrendako lehen heuristikoak erauzten ditu hautagai gehien eta azkenak gutxien, bi horiek baitira, hurrenez hurren, hautagaiak erauzteko murriztapen gutxien eta gehien kontuan hartzen dituztenak: zenbat eta murriztapen gehiago, orduan eta hautagai gutxiago erauzten dira, eta alderantziz. Arrazoi berberagatik, lehenak

[6]Irudiko dependentzia sintaktikoak Dependentzia Unibertsalen ereduaren arabera daude zehaztuta, eta laburtzapenak ere eredu horretan erabiltzen direnak dira (Aranzabe *et al.*, 2019): *dobj*, objektu zuzena; *nsubj*, izen-subjektua; *conj*, juntagailua; *cc*, koordinazio-juntagailua; *acl* adjektibo-perpausa.

"okerreko" hautagai gehiago erauzten ditu azkenak baino, benetan agerpen literalak ez diren hautagai gehiago alegia. Dena dela, laurak hartu ditugu guk kontuan, gure helburua ez baitzen erauzketa ahalik eta doiena egitea baizik eta ahalik eta zabalena, eskuzko azterketa ere ahalik eta osoena izan zedin.

4.2 Hautagaien sailkapena

Corpusean etiketatutako UF guztiak erauzi, eta zerrenda bat osatu dugu haiekin: testuingururik gabeko hitz-konbinazioen multzo bat. Hemendik aurrera, *jatorrizko UF* deituko diegu zerrenda horretako hitz-konbinazioei. Jatorrizko UFen barruko lemak corpusean elkarrekin agertzen direnean, berriz, *agerpenez* hitz egingo dugu, eta hiru multzotan sailkatuko ditugu: agerpen idiomatikoak, agerpen literalak eta kointzidentziazko agerpenak. Izan ere, lan honen aztergai nagusia agerpen literalak badira ere, beste bi kontzeptuak ere kontuan hartzekoak dira, zenbaitetan ez baita oso argi egoten non dagoen multzo baten eta besteen arteko muga. Saia gaitezen, beraz, lan honen ardatz diren hiru kontzeptuak argitzen.

UF baten lemak corpusean elkarrekin agertzen direnean, corpuseko hitz-konbinazio hori **agerpen idiomatiko**tzat (AI) jotzen da, baldin eta (1) lemen arteko erlazio sintaktikoa bat badator jatorrizko UFaren egitura sintaktikoarekin[7], eta (2) esanahi idiomatikoa badu. Esate baterako, *adarra jo* hitz-konbinazioa aintzat hartuta eta jatorrizko UFan *adar* izena *jo* aditzaren objektu zuzena dela jakinik, 11. adibideko agerpena idiomatikoa litzateke:

(11) *Ez egin jaramonik,* ***adarra jotzen*** *ari zaizu eta.* → egitura sintaktiko berbera, esanahi idiomatikoa

Aldiz, corpuseko hitz-konbinazio bat UF baten **agerpen literal**tzat (AL) jotzen dugu lemen arteko erlazioa bat badator jatorrizko UFko lemen artekoarekin baina agerpenaren esanahia idiomatikoa ez bada. Aurreko adibide berberarekin jarraituz, honako hauek *adarra jo* UFaren agerpen literalak lirateke:

(12) *Pinaburuak zuhaitzaren adarrak jo zituen erortzean.* → egitura sintaktiko berbera, esanahi literala

(13) *Beleak zuhaitzaren bi adar jo zituen.* → egitura sintaktiko berbera, esanahi literala

Azkenik, hitz-konbinazio bati **kointzidentziazko agerpen** (KA) deritzogu, baldin eta lemen arteko erlazio sintaktikoa desberdina bada jatorrizko UFaren egitura sintaktikoaren aldean. Adibidez, irudiko (b) eta (d) esaldiek *arrastoa utzi* UFaren kointzidentziazko agerpen bana biltzen dute, eta beste honek, berriz, *adarra jo* UFaren bat:

(14) *# Zuhaitzaren adar bat hautsi zuen baloia jo nahian.* → egitura sintaktiko desberdina

Etiketatze-lana egiteko, hizkuntza bakoitzeko etiketatzaileei taula banatan gorde zaizkie honako datuak: jatorrizko UFa eta hari dagokion etiketa (LVC ala VID), hautagai bakoitzari dagokion esaldia osorik, hautagaiaren barruko lemen kategoria gramatikalak, eta zer heuristikok erauzi du(t)en hautagai bakoitza. Hizkuntza bakoitzeko hiztun aditu banak hartu du parte etiketatze-lanean, portugesezko zatian salbu, horretan bi etiketatzailek jardun baitute.

Aintzat harturik PARSEMEren irizpideekin bat datozen agerpen idiomatikoak etiketatuta daudela corpusean, heuristikoek erauzitako hautagaiak kointzidentziazkoak ala literalak ziren esan dute etiketatzaileek[8]. Horrez gain, agerpen literalak hiru azpimultzotan sailkatu dituzte, agerpen idiomatikoetatik bereizteko kontuan hartu beharreko informazioaren arabera.

- LITERAL-MORPH: agerpen idiomatikoetatik bereiz daiteke murriztapen morfologikoak kontuan hartuta. Esate baterako, 12. adibidea multzo honetan sailkatzekoa da, *adarra jo* UFa ez baita inoiz pluralean erabiltzen idiomatikoa denean.
- LITERAL-SYNT: agerpen idiomatikoetatik bereiz daiteke murriztapen sintaktikoak kontuan hartuta. Adibidez, gorago eman dugun 13. adibidea *adarra jo* UFtik bereizteko, nahikoa da izen-sintagmako *zuhaitzaren* modifikatzaileari eta *bi* determinatzaileari begiratzea, jatorrizko UFak ez baitu halakorik onartzen[9].

[7] Aurreko atalean azaldu bezala (3.1), egitura sintaktiko baliokidetzat jotzen ditugu forma kanonikora ekarrita egitura berbera dutenak.

[8] Ohar bedi agerpen idiomatiko guzti-guztiak ez daudela benetan etiketatuta corpusean, eskuzko etiketatze-lanean akatsak ere egoten baitira. Azterketa egitean multzo gehiago sortu ditugu erroreetatik eratorritako hautagaiak sailkatzeko, baina hemen, laburtze aldera, kointzidentziazko agerpenei eta literalei dagozkienak bakarrik aipatuko ditugu.

[9] Halako kasuak morfologiaren eta sintaxiaren arteko mugakoak diren arren, ataza honetan Dependentzia Unibertsalaren ildoari jarraitu diogu, eta, hortaz, fenomeno morfologikotzat hartu ditugu hitz baten barruan gertatzen diren aldaketa guztiak (*adarra, adarretik, adarrek...*), eta sintaktikotzat hitzetik kanpo gertatzen direnak (*adar bat, adar luzeak...*).

- LITERAL-OTHER: murriztapen morfosintaktikoak ez dira nahikoa agerpen idiomatikoetatik bereizteko; testuinguruari, semantikari edo hizkuntzaz kanpoko ezaugarriei begiratu beharra dago horretarako. Horixe gertatzen da, adibidez, 1. irudiko (a) eta (c) esaldiekin, ezin baitira horko *arrastoa utzi* literalak agerpen idiomatikoetatik bereizi morfosintaxiari begiratuta bakarrik.

5 Emaitzak

Etiketatze-lanaren estatistika nagusiak 2. taulan jaso ditugu. Oro har, idiomatikotasun-tasa[10] oso altua da, eta emaitzak oso antzekoak dira hizkuntza guztietan: % 96tik % 98ra bitartekoak. Hortaz, gure hipotesi nagusia betetzen da: UFen agerpen literalak oso urriak dira testu errealetan[11].

	DE	EL	EU	PL	PT
Corpuseko UFak	3.823	2.405	3.823	4.843	5.536
Hautagaiak	926	451	2.618	332	1.997
Kointzidentziazkoak	**% 2,6** (24)	**% 27,9** (126)	**% 42,4** (1110)	**% 61,1** (203)	**% 33,5** (668)
Literalak	**% 8,5** (79)	**% 11,5** (52)	**% 3,5** (91)	**% 29,5** (98)	**% 12,9** (258)
↪ literal-morph	% 0,8 (7)	% 5,5 (25)	% 1,9 (51)	% 1,2 (4)	% 3,7 (73)
↪ literal-synt	% 1,5 (14)	% 2 (9)	% 0,7 (19)	% 8,1 (27)	% 2,2 (44)
↪ literal-other	% 6,3 (58)	% 4 (18)	% 0,8 (21)	% 20,2 (67)	% 7,1 (141)
Bestelakoak	% 88,9 (823)	% 60,5 (273)	% 54,1 (1417)	% 9,3 (31)	% 53,6 (1071)
Idiomatikotasun-tasa	**% 98**	**% 98**	**% 98**	**% 98**	**% 96**

2. Taula: Etiketatze-lanaren estatistika orokorrak, hizkuntza guztietan.

Arrazoi tipologikoak direla medio, hautagaien kopuruan alde nabarmena dago hizkuntza batetik bestera: mutur batean poloniera dago, heuristikoek 332 hautagai erauzi baitituzte 4.843 UF etiketatatik abiatuta, eta beste muturrean, berriz, euskara, 2.618 hautagai izan baitira 3.823 UF etiketaren bidez erauzitakoak. Kointzidentziazko agerpenen ugaritasuna da, batez ere, alde hori markatzen duena, eta datorren azpiatalean azalduko dugu ugaritasun horren zergatia. Nolanahi ere, taulak argi erakusten du morfosintaxiaren bidez ebatzi ezin diren kasuak oso gutxi direla, euskaraz bereziki. Hain zuzen, multzo horrek ez du hautagai guztien % 1 ere osatzen. Gure bigarren hipotesia ere betetzen da, beraz: morfosintaxiak garrantzi handia du idiomatiko-literal bereizketan, kointzidentziazko agerpenak sintaxiaren bidez ebatz baitaitezke, eta literal gehien-gehienak ere morfologiaren edo sintaxiaren bidez.

Behin emaitza orokorrak ikusita, azal dezagun oro har nolakoak diren multzo bakoitzean sailkatutako hautagaiak, eta ikus dezagun zer berezitasun dituen euskarak beste hizkuntzen aldean.

Kointzidentziazko agerpenen ezaugarriak

Euskara da, alde handiz, kointzidentziazko agerpen gehien dituen hizkuntza (2. taula). Agerpen horietako askok eta askok postposiziodun izenak dituzte barnean, jatorrizko UFan izenak halakorik ez bazuen ere. Izan ere, heuristikoek lemari begiratuta bakarrik erauzten dituzte hautagaiak, eta, postposizio- eta kasu-markak lematik kanpo geratzen direnez, izenari eransten zaizkion marka horiek ez dira kontuan hartzen. Hori gertatzen da, esate baterako, ondorengo adibideetan. *Aurre egin* UFa kontuan hartuta, *aurre* eta *egin* lemak bilatu dira, eta hautagaitzat erauzi dira 16. eta 17. adibideetako hitz-konbinazioak.

(15) *Arazoei **aurre egin** zien.*

(16) *# Donostiako udaletxearen aurrean egin dute elkarretaratzea.*

(17) *# Hitz egiten hasi aurretik egin beharrekoak.*

Izenari postposizioa gehitzeak osagaien arteko erlazio sintaktikoa aldatzen du adibide horretan, eta, hortaz, hautagaia kointzidentziazko agerpentzat sailkatu beharrekoa da. Etiketa bestelakoa litzateke jatorrizko UFak postposizioren bat barne hartuko balu (adib.: *aurrera egin*), litekeena baita halakoetan erlazio sintaktikoa berbera izatea hautagaiak postposizio desberdina izanik ere, eta LITERAL-MORPH etiketa beharko bailuke kasu horietan.

[10] Idiomatikotasun-tasa = AI/(AI+AL).

[11] Lehenago ere esan dugunez (8. oin-oharra), azterketa osoan etiketa gehiago erabili ditugu (Savary *et al.*, 2019), baina lan honetan, datuak laburtzeko asmoz, *Bestelakoak* multzoan bildu ditugu corpuseko erroreei dagozkienak, analizatzailearen erroreei dagozkienak eta testuinguru murritzegia dutenak ondo sailkatu ahal izateko.

Bistan denez, heuristikoak ez dira hizkuntza aglutinatiboetan pentsatuz sortu hasiera batean. Ahalik eta heuristikorik orokorrenak sortu nahi izan direnez, lemari bakarrik begiratu zaio, baina horrek alferreko hautagai asko eta asko erauzarazi ditu euskaraz, lan honetako hizkuntza aglutinatibo bakarrean. Etorkizuneko lanei begira, hobe litzateke postposizio-markak ere lexikalizatutzat hartzea beti, gainerako hizkuntzetan preposizio lexikalizatuak (hitz beregainak izanik) kontuan hartzen diren modu berean. Hartara, *atzera egin* UFtik ez lirateke *atzean egin* eta halako hautagaiak eratorriko, ingelesezko *be in love* (lit. *maitasunean egon* 'maiteminduta egon') UFtik *be of love* (lit. *maitasuneko egon*) eratorriko ez litzatekeen bezalaxe.

Agerpen literalen ezaugarriak

Agerpen literalen ezaugarriak desberdin samarrak dira UF mota batetik bestera. Esate baterako, idiomatikotasuntasa askoz ere altuagoa da LVCetan VIDetan baino, edozein hizkuntzaz ari garela ere. Izan ere, LVCak nahiko konposizionalak dira semantikari dagokionez, izenak bere ohiko esanahia gordetzen baitu eta aditzak ezaugarri morfologikoak baino ez baitizkio gehitzen normalean (3.2. atala). Intuizioz ere, ez da hain erraza multzo horretako UFek agerpen literalak dituztela pentsatzea. Dena dela, badaude halako kasu batzuk, non LVC barruko bi lemak elkarrekin agertzen baitira baina ez baitituzte LVCaren ezaugarriak betetzen. Hori gertatu ohi da, adibidez, izen batek esanahi predikatiboa eta ez-predikatiboa izan dezakeenean, alegia, batzuetan bakarrik egiten dionean erreferentzia ekintza edo egoera bati (18–19. adibideak).

(18) *Sekulako **laguntza eman** dit kirolean eta kiroletik kanpo.*

(19) *Enpresa berriak sustatzeko laguntzak emango ditu Udalak.*

Goiko adibideetako lehenengoan, *laguntza eman* idiomatikoki dago erabilita, baina ez bigarrenean PARSEMEren gidalerroen arabera, diru-laguntza bati buruz ari baita eta ez laguntzeko ekintzari buruz. Esanahi idiomatikoan *laguntza* beti singularrean erabiltzen denez eta 19. adibidean pluralean dagoenez, bigarren adibide horri LITERAL-MORPH etiketa eman diogu.

Bestalde, VID motako UF asko metaforetatik datoz, eta intuizioz errazagoa da halakoek esanahi figuratiboa eta literala dutela pentsatzea. Era horretakoa da, adibidez, *atzera bota* UFa; 20. esaldian idiomatikoki erabilita dago, eta 21.ean, aldiz, literalki, zerbait fisikoki atzerantz botatzea adierazten baitu. Esaldi horri LITERAL-OTHER etiketa dagokio.

(20) *Irakasleen eskaerak **atzera bota** ditu Hezkuntzak.*

(21) *Pase aparta eman, eta baloia atzera bota dio taldekideari.*

Uste genuen bezala, eta 2. taulak agerian uzten duenez, VIDetako asko bereiz daitezke morfosintaxiari begiratuta, euskaraz bereziki. Esate baterako, *gai izan* UFan izena ez da inoiz adjektibo batez lagunduta egoten, eta ezaugarri horretxeren bidez jakin liteke 23. adibideko agerpena literala dela, LITERAL-SYNT motakoa zehazki.

(22) *Lau langiletik bat **gai da** euskaraz aritzeko.*

(23) *Horixe da gaurko gai nagusia.*

6 Ondorioak eta etorkizuneko lanak

Fraseologia mailan etiketatutako corpus bat oinarritzat hartuta, Unitate Fraseologikoen agerpen literalez jardun dugu lan honetan, eta erakutsi dugu halako agerpenak *urre baina urri* direla. Izan ere, *urre* dira batetik, Hizkuntzaren Prozesamenduko tresnek behar-beharrezkoa dutelako esanahi idiomatikoak eta literalak bereiztea, tresna linguistikoek taxuzko emaitzak sortuko badituzte. Baina bestetik, *urri* dira, corpusak erakusten baitu oso gutxitan erabiltzen direla literalki praktikan. Horrez gain, gure lanaren arabera, agerpen literal gehienak (bai eta kointzidentziazko agerpenak ere) ezaugarri morfosintaktikoei begiratuta bereiz daitezke agerpen idiomatikoetatik, eta datu horrek garrantzi handia du HPko tresnen garapenerako.

Aurrera begira, interesgarria litzateke azterketa hau beste corpus mota batzuetan berregitea, UF asko testu mota jakinetan bakarrik agertzen baitira, eta horrek bide emango bailiguke testu genero batetik bestera zer-nolako aldea dagoen ikusteko. Horrez gain, hizkuntza gehiago kontuan hartzea ere lan hau zabaltzeko beste modu bat da; hala ikusiko genuke ondorioak antzekoak ote diren beste hizkuntza batzuetan ere.

Erreferentziak

Abeillé, Anne, eta Yves Schabes. 1989. Parsing idioms in lexicalized tags. In *Proceedings of the fourth conference on European chapter of the Association for Computational Linguistics*, 1–9. Association for Computational Linguistics.

Aranzabe, Maria Jesus, Aitziber Atutxa, Kepa Bengoetxea, Arantza Díaz de Ilarraza, Iakes Goenaga, Koldo Gojenola, eta Larraitz Uria. 2019. Dependentzia unibertsalen eredura egokitutako euskarazko zuhaitz-bankua. *EKAIA, EHUko zientzia eta teknologia aldizkaria* .

Azkarate, Miren, 1987. *Hitz elkartuak euskaraz*. Deustuko Unibertsitatea tesia.

Baldwin, Timothy, eta SuÑam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing*, ed. by Nitin Indurkhya eta Fred J. Damerau, 267–292. Boca Raton, FL, USA: CRC Press, Taylor and Francis Group, 2 edition.

Cacciari, Cristina, eta Paola Corradini. 2015. Literal analysis and idiom retrieval in ambiguous idioms processing: A reading-time study. *Journal of Cognitive Psychology* 27.797–811.

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, eta Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics* .

Geeraert, Kristina, R Harald Baayen, eta John Newman. 2018. "spilling the bag" on idiomatic variation. *Multiword expressions at length and in depth* 1–33.

Gurrutxaga, Antton, Iñaki Alegria, eta Xabier Artola. 2014. Idiomatikotasunaren karakterizazio automatikoa: izena+aditza konbinazioak. *Ekaia. EHUko Zientzia eta Teknologia aldizkaria* .

Hualde, José Ignacio, eta Jon Ortiz de Urbina. 2003. *A grammar of Basque*. Walter de Gruyter.

Inurrieta, Uxoa, Itziar Aduriz, Arantza Diaz de Ilarraza, Gorka Labaka, eta Kepa Sarasola. 2017. Aditza+izena konbinazioen itzulpen automatikoa, arau linguistikoen bidez. In *IkerGazte: nazioarteko ikerketa euskaraz. Kongresuko artikulu-bilduma. Giza zientziak eta artea*, 158–166.

——, Itziar Aduriz, Ainara Estarrona, Itziar Gonzalez-Dios, Antton Gurrutxaga, Ruben Urizar, eta Inaki Alegria. 2018. Verbal multiword expressions in basque corpora. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 86–95.

Savary, Agata, Marie Candito, V Barbu Mititelu, Eduard Bejček, Fabienne Cap, eta M van Gompel. 2018. Parseme multilingual corpus of verbal multiword expressions. 87–147.

——, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Inurrieta, Voula Giouli, eta Ivelina Stoyanova. 2019. Literal occurrences of multiword expressions: rare birds that cause a stir. argitaratze–bidean.

——, eta Silvio Ricardo Cordeiro. 2017. Literal readings of multiword expressions: as scarce as hen's teeth. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, 64–72.

——, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, eta others. 2015. Parseme–parsing and multiword expressions within a european multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*.

Sheinfux, Livnat Herzig, Tali Arad Greshler, Nurit Melnik, eta Shuly Wintner. 2017. Verbal mwes: Idiomaticity and flexibility. *Representation and Parsing of Multiword Expressions* 5–38.

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

Waszczuk, Jakub, Agata Savary, eta Yannick Parmentier. 2016. Promoting multiword expressions in a*tag parsing. In *26th International Conference on Computational Linguistics (COLING 2016)*.

Zabala, Igone. 2004. Los predicados complejos en vasco. In *Las fronteras de la composición en lenguas románicas y en vasco*, 445–534. Deustuko Unibertsitatea.

7 Eskerrak eta oharrak

Artikulu hau lankidetzan eginiko lan zabalago baten parte da. Azterketa osoaren xehetasunak *Literal occurrences of multiword expressions: rare birds that cause a stir* artikuluan daude jasota (Savary *et al.*, 2019). Lankidetza hori PARSEME proiektutik sortu da (IC1207 COST Action), eta honako egile hauek ere parte hartu dute: Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch eta Voula Giouli.