# Fraseologia konputazionalaren inguruan bueltaka:
## tesigile baten atzerabegirakoa

Uxoa Iñurrieta (HiTZ zentroa, Ixa taldea, UPV/EHU)

Artikulu hau idazten hasi aurretik, nire doktoretza-tesiaren laburpen moduko bat egiteko asmoa nuen, amaitu berri dudan ikerketa- lanaren gakoak orrialde gutxi batzuetara ekartzekoa. Baina, gai berari beti ikuspuntu beretik begiratzeaz nazkatuta-edo, ia nahi gabe, ihes egin diot artikulu akademikoen zurruntasunari; ikerketa-lanetan erabili ohi diren atalkatzeak alde batera utzi (sarrera, arloaren egoera, metodologia, emaitzak…), eta beste era batera antolatu dut lan hau. Gaiaren akademikoan, artikulu ez-akademiko bat atera zait azkenean, eta hari nagusia, doktoretza-tesiak markatu beharrean, nire tesigintza-ibilbideak markatu du. Tesia hasi eta amaitu bitarteko urteen kontakizun kronologikoa da hau, beraz, eta lehen pertsonan kontatua.

### Lehen harria ekarri genuenekoa

Artean Hizkuntzaren Azterketa eta Prozesamendua masterra egiten ari nintzela, Ixa taldean hasi nintzen praktiketan, Euskal Herriko Unibertsitateko Informatika Fakultatean, masterra antolatzen zuen ikerketa-taldean bertan. Ezer gutxi nekien hizkuntzalaritza konputazionalaz orduan, eta burua bete-bete bueltatzen nintzen etxera, bai masterreko eskoletatik eta bai lanetik. Informazio-jasa ikaragarria zen, eta aitortu behar dut, hain ikasturte trinkoan, informatika-kontu haietatik guztietatik zati handi samar batek egiten zidala ihes. Hala ere, mundu berri bat zabaldu zitzaidan urte hartan: hizkuntza prozesatzeko tresna informatikoen tripak ezagutzen

hasi nintzen pixkanaka, eta, horren aitzakian, hizkuntzari berari ere ordura arte erreparatu ez nion bezala erreparatzen.

Praktiketan hasi eta denbora gutxira eskaini zidaten doktoretza-tesia egiteko aukera. Nik banuen interesa, eta banekien gutxi gorabehera zeren inguruan ikertu nahi nuen: itzulpena eta fraseologia interesatzen zitzaizkidan. Tesi baten gaia aukeratzea, ordea, ez da hain sinplea. Ikerketaren mundua ia batere ezagutzen ez duenari bururatzen zaizkion ideiak, normalean, behar baino askoz ere zabalagoak izaten dira tesi baterako, eta zedarriak jarri behar izaten zaizkie: pixka bat mugatu, eta beste pixka bat, eta beste pixka bat… Tesi-proiektu bideragarri samar bat osatu arte. (Hasierako proiektuak eta bukaerako emaitzak ere ez dute zerikusi handirik izaten, baina hori beste kontu bat da.)

Zedarritze-lan horretan, ezinbestekoa da eskarmentu handiagokoen laguntza, eta hor hasten da tesi-zuzendarien lana. Orduan, Arantza Diaz de Ilarrazak koordinatzen zuen Ixa taldearen jarduna, eta hark egin zidan proposamena: zuzendarietako bat Itziar Aduriz izango zen, hizkuntzalaria, eta, alderdi informatikoari zegokionez, berriz, Kepa Sarasola eta Gorka Labaka izango nituen gidari. Ofizialki, Aduriz eta Labaka agertu dira zuzendari tesian, baina Diaz de Ilarrazak eta Sarasolak ere gurekin batera egin dute bide osoa.

Labur esateko, honako hau zen doktoretza-tesiaren motibazioa. Demagun beheko hiru esaldiok euskaratu behar ditugula:

(1) No podemos *correr* ese *riesgo*.
(2) *Metían* muchísimo *ruido*.
(3) La pareja *contrajo matrimonio*.

Gaztelaniaz eta euskaraz ondo egiten dugunontzat, esaldi horiek itzultzeak ez du aparteko zailtasunik: inork ez luke jarriko lehen esaldiaren itzulpenean *korrika egin*ik, ez bigarrenean *sartu*rik, ez eta hirugarrenean *uzkurtu*rik ere. Orduko itzultzaile automatikoek, ordea, uste zuten arriskuak *korrika egiten* zirela, zarata *sartu* egiten zela eta ezkontzak *uzkurtu* egiten zirela. Bai, badakit, oraingo sistemak beste kontu bat dira, baina 2014an artean urrun ikusten zen itzultzaile automatikoek gaur egun duten kalitatea izatea.

Bada, halako itzulpen traketsak nola konpondu aztertzea zen gure egitekoa. Zehazki, aditza+izena motako Unitate Fraseologikoak (UFak) izango genituen aztergai, kolokazioak eta lokuzioak, gure aurrekari nagusien arabera horiek baitziren zailenetakoak automatikoki prozesatzeko (Urizar, 2012; Gurrutxaga, 2015). Ez zen nolanahiko lana:

- Nola lortu ordenagailuek UFak automatikoki identifikatzea? *Liburua irakurri* eta *hanka sartu*, biak dira izena+aditza konbinazioak, baina lehena konbinazio librea da eta bigarrena ez.
- Nola bereizi automatikoki zein UF itzultzen diren hitzez hitz eta zein ez? Gaztelaniazko *meter la pata*ren ordaina emateko, erabil daiteke *hanka sartu* euskaraz, baina *estirar la pata*ren zentzu idiomatikoa (hiltzea, alegia) ezin adieraz daiteke *hanka luzatu* esanda.

- Eta nola lortu hitzez hitz itzuli ezin diren UF horien ordain egokia ematea? Hots, nola adierazi ordenagailuari honako hau, adibidez: *tomar el pelo*rentzat, *adarra jo* da euskarazko ordain egokia, ez *ile* eta ez *hartu*; aldiz, *meter ruido*ren kasuan, izenari ohiko ordaina ematen zaio, *zarata*, baina aditzari, *egin* edo, bestela, espero denaren justu kontrakoa, *sartu* ordez *atera*.

Proiektu-proposamena idatzi, eta 2014ko otsailean ekin genion ikerketa-lanari, Ekonomia eta Lehiakortasun Ministerioaren doktoretza-aurreko dirulaguntza bat lortuta. Ekarria genuen lehen harria.

## Aparejurik gabe arrantzan hasi ginenekoa

Kanpotik begiratuta, badirudi doktoretza-tesi batean gaiaren inguruan idatzi den guzti-guztia irakurri beharra dagoela beste ezertan hasi aurretik. Eta bai, jakina, hori litzateke ikerketa egiteko modurik onena… baldin eta denbora-mugarik ez bagenu. Benetan, berriz, lan jakin batzuk bakarrik irakurri ahal izaten dira hasierako hilabeteetan, eta, behin arloan egin den ikerketaren ikuspegi orokor samar bat lortutakoan, norberaren lanari heldu behar izaten zaio. Aparejurik gabe arrantzan hastearen pare ia-ia, bibliografiaren zati handi bat ezagutu gabe tresna asko falta izaten baitira lana taxuz egiteko.

Ekin genion, nolanahi ere, eta lehenengo lana hiztegi elebidunen gainean egitea erabaki genuen. Izan ere, abiapuntu gisa erabiliko genuen itzultzaile automatikoa *Matxin* zen (Mayor *et al.*, 2009), eta sistema horrek gramatika-arauak eta hiztegi elebidunak erabiltzen zituen itzulpenak egiteko. *Elhuyar* gaztelania-euskara eta euskara-gaztelania hiztegia hautatu genuen lehen azterketa horretarako, eta aditza+izena (eta izena+aditza) motako hiztegi-sarreren eta ordainen analisi xehe-xehea egin genuen, batez ere morfosintaxiari eta lexikoari begiratuta.

Azterketa horren bidez eginiko ekarpenen artean, besteak beste, zenbakitan jarri genuen lehenagotik genuen ustea, hau da, aditza+izena motako UF asko eta asko direla hitzez hitz itzulezinak. Hain zuzen ere, gaztelaniazko aditza+izena motako hiztegi-sarreren % 11k eta euskarazkoen % 7k bakarrik dute hitzez hitzeko ordaina *Elhuyar* hiztegian. Eta, gainera, lexikoa alde batera utzita ere, gaztelaniazko sarreren erdiak eta euskarazkoen herenak baino gutxiagok dute aditz batez eta izen batez osatutako ordainen bat; esan nahi baita, gainerako guztiek bestelako osaera morfosintaktikoko ordainak dituztela:

(4) *alzar (el) vuelo* → *hegan hasi* (ES, aditza+izena; EU, adberbioa+aditza)
(5) *hanka egin* → *largarse* (EU, izena+aditza; ES, aditza)

Bi argitalpenetan eman genuen azterketa horren berri: *Linguamática* aldizkarian euskaraz (Iñurrieta *et al.*, 2014), eta John Benjamins argitaletxearen *Multiword Units in Machine*

*Translation and Translation Technologies* liburuko kapitulu batean ingelesez (Iñurrieta *et al.*, 2018a).[1]

Lan hori egin bitartean, bibliografian sakontzen ere jarraitu genuen, azterketa linguistikoez gain hizkuntzalaritza konputazionalari buruzko lan aplikatuagoak irakurtzen batez ere; aparejuak lortzen, azken batean, gure lehen lan esperimentalari ekin nahi baikenion hurrena.

Bide horretan, giltzarria izan zen niretzat PARSEME[2] Europako proiektua ezagutzea (Savary *et al.*, 2015), fraseologia konputazionaleko ikerlarien artean elkarlana sustatzea baitzen egitasmo horren helburua. Prestakuntza-saio, bilera eta lantegi ugari antolatu zituzten, eta, zuzendariek animatuta, aukera izan nuen haietako batzuetan parte hartzeko; hala ezagutu nituen nire tesian aztarna nabarmena utzi duten zenbait ikerlari eta erreferentzia garrantzitsu. Astebeteko prestakuntza-saio batera joan nintzen lehenik, Pragara, eta, ni proiektuan sartzearekin batera, euskara ere batu zen proiektuko hizkuntzen zakura.

## Gora eta behera ibili ginenekoa

Ez dakit tesigile guztiei gertatuko zaien gauza bera, baina ni, hasierako artikulu horiek argitaratu berri eta bibliografia irakurri eta irakurri, oso baikor nengoen lehen urte eta pikoan: gaian gero eta jantziago sentitzen nintzen, eta oraindik banuen astia tesian ia nahi nuena egiteko. Alta, jendeak ez du alferrik esaten doktoretza-tesiak errusiar mendiak bezalakoak direla. Besteak beste, zenbat eta gehiago sakondu gai batean, orduan eta agerikoagoa egiten zaiolako bati zer ez dakien, zer ez duen egin eta zertan erratu den, eta horrek nahigabea sortzen du.

Laster jarriko zen nire baikortasuna kolokan. Baina gatozen, momentuz, ikerketa-lanaren muinera berriro. Hasiak ginen lehen lan esperimentalean, eta kontua honako hau zen: erakutsi nahi genuen UF askok izaten dituztela ezaugarri morfosintaktiko oso markatuak eta, ezaugarri horiek aintzat hartuz hain zuzen posible dela UFen prozesamendua hobetzea.

Har ditzagun, adibidez, gaztelaniazko eta ingelesezko bi UF hauek: *tomar el pelo* (lit. *ilea hartu*, 'adarra jo') eta *be in love* (lit. *maitasunean egon,* 'maiteminduta egon'). Gaztelaniazkoa beti erabiltzen da artikulu mugatuarekin eta singularrean, eta ingelesezkoa, berriz, beti singularrean baina determinatzailerik gabe. Gainera, gaztelaniazkoan, izen-sintagma aditzaren objektu zuzena da, eta ingelesezkoan, preposizio-sintagma, aditzaren osagarri zirkunstantziala. Horiei erreparatuta, eta analizatzaile morfosintaktiko automatikoak erabiliz, erraza litzateke beheko esaldi hauetan UFen agerpenak (6a eta 7a) eta bestelakoak (6b-c eta 7b-c) bereiztea, nahiz eta UFen barruko osagai lexikoak esaldi guztietan agertu.

(6) a. Te está *tomando el pelo*.

1. Bigarren artikulu hori, 2018an argitaratu bazen ere, 2015ean bidali eta onartu zen.
2. PARSEME: PARSing and Multiword Expressions. Towards linguistic precision and computational efficiency in natural language processing.

b. Los expertos lo comprobaron tras *tomar* un *pelo* rubio y aplicarle spray de grafeno con agua.
c. Debemos *tomar* muestras de *pelo* y fibra.

(7) a. They *are in love*.
b. They *are in* the *love* of God.
c. They may *be in Loves* Park or nearby.

Gaztelaniazko 117 eta ingelesezko 173 UFrekin egin genuen proba: haietako bakoitzaren ezaugarri bereizgarriak zein ziren eskuz aztertu, eta esperimentu bat egin genuen, ikusteko ea informazio hori guk uste bezain baliagarria zen identifikazio-lana hobetzeko. Oro har, emaitza onak lortu genituen, gai baikinen askoz ere agerpen gehiago identifikatzeko *Matxin*ek baino.

Lan horren zati handi bat Sussexeko unibertsitatean egin nuen, Ingalaterran, John A. Carroll irakaslearen gidaritzapean. Hilabete interesgarriak izan ziren: asko ikasi nuen informatika-kontuez eta ikerketa-metodologiaz, eta, gainera, artikulu bat onartu ziguten arloko kongresu garrantzitsu batean: COLINGen[3] (Iñurrieta *et al.,* 2016b). Gure lantxoak bere txokoa izan zuen Osakan egin zen kongresu hartan, Japonian, beste hirurehun bat lanen artean.

Ordurako iritsia zen sare neuronalen[4] oldea hizkuntzalaritza konputazionalera, eta nabarmena zen halako ikerketa-lanek bereganatzen zutela parte-hartzaileen interes gehien. Euskal Herrian ere, orduantxe ari ziren Ixako taldekide batzuk, beste zenbait erakunderekin batera, MODELA itzultzaile automatikoa sortzen (Etchegoyhen *et al.*, 2018), euskal itzulpengintza automatikoaren historian lehena eta geroa bereiziko zituen sistema.

Testuinguru horretan hasi zen nire baitako errusiar mendia forma hartzen, eta esango nuke ordutik aurrerako beheraldi gehienek sare neuronalekin izan zutela lotura. Izan ere, itzultzaile neuronalek corpusetatik ateratzen dute itzulpenak egiteko informazioa, eta hitzen arteko agerkidetza da, hain zuzen, sistema horiek entrenatzeko erabiltzen den irizpide nagusia. UFen ezaugarri bereizgarrietako bat osagai-hitzek elkarrekin agertzeko duten joera izanik, ez da harritzekoa, adibidez, MODELAren ondotik sorturiko *Itzultzailea.eus*-ek honela itzultzea artikulu honetako lehen hiru adibideak:

(8) Ezin dugu *arrisku* hori *hartu*.
(9) *Zarata* handia *egiten* dute.
(10) Bikotea *ezkondu* egin zen.

Gauzak hala, non geratzen zen gure lana? Tira, egia da hor jarritako adibideak sinple samarrak direla eta, itzulgaia zein den, izaten dela zer zuzendua fraseologiari dagokionez ere itzulpen automatikoetan. Baina nik dagoeneko ez nuen lehen bezain argi ikusten tesi-lanaren ekarpena,

3. International Conference on Computational Linguistics

4. Adimen artifizialean, sare neuronaletan oinarritutako sistemek giza garuna imitatzea dute helburu. Azken urteotan, teknika hori nagusitu da hizkuntzalaritza konputazionalean eta, zehazki, itzulpen automatikoan.

eta etengabe galdetzen nion neure buruari ez ote zen hobe ordura artekoak alde batera utzi eta beste zerbaitetan hastea. Ez nuen halakorik egin, ordea, eta, ilusioa apalduta ere, lanarekin jarraitzea erabaki nuen.

UFak identifikatzeko eginiko esperimentuaren ondoren, identifikatutako UFei ordaina emateko proposamena egin genuen (Iñurrieta *et al.,* 2016a), eta IkerGazte kongresuan eta EACL[5] kongresuko lantegi batean aurkeztu (Iñurrieta *et al.*, 2017), Iruñean eta Valentzian. Ez genuen emaitza txundigarririk lortu, baina bai *Matxin*en jatorrizko sistemarenak baino hobeak; hiru ebaluatzaileren arabera, UFen itzulpenen % 76 hobeak ziren gure proposamena aplikatu ondoren.

Beraz, emaitzak ez ziren txarrak berez, ez identifikazioari zegokionez eta ez itzulpenari zegokionez ere. Baina pentsa: urtebete inguru behar izan genuen lan hori egiteko, eta UF eta ordain gutxi batzuk bakarrik aztertu ahal izan genituen ordura arte. Eginikoaren dimentsioa argi samar ikusten da gaztelaniazko corpus anotatu[6] batean bilaketak eginda: landutako UFekin bakarrik, corpuseko UF guztien % 8 besterik ezingo genuke hauteman. Hortaz, tesi-lana biribiltzeko, eskalabilitate-arazo horri aurre egitea falta zitzaigun: bideren bat bilatu beharra genuen eginiko azterketa linguistikoa automatizatzeko eta, hala, eskala handiagoan aplikatu ahal izateko.

Horretan hasiak ginela, Ingeniaritza Eskolan euskara teknikoko eskolak emateko aukera egokitu zitzaidan, eta, dirulaguntza amaitu baino hilabete batzuk lehenago, jardunaldi osoan ikerketan aritzeari utzi nion. Eskolak eta ikerketa uztartu beharrak tesiaren erritmoa motelarazi zidan hein batean, baina ez zidan erabat galarazi lanean jarraitzea. Are gehiago: esango nuke lagungarria ere izan zela, burua tarteka tesitik urruntzeak perspektiba pixka bat harrarazi zidalako eta garbiago ikusi nituelako, tesi-lanaren ahulguneak ez ezik, indarguneak ere.

Azterketa linguistikoa automatizatzeko, hainbat pausotan banatu genuen gure proposamena: (1) UF-zerrendak lortu genituen hiztegietatik, eta UFen osagai-hitzen bilaketak egin genituen corpus elebakar batean, eskuzko azterketan ezarritako zenbait irizpide morfosintaktikoren arabera; (2) lortutako datu horiekin guztiekin, estatistikari erreparatuz batez ere, UFak ezaugarri morfosintaktikoka multzokatu genituen; (3) corpus paraleloetan, gaztelaniazko UFei euskarazko zer ordain ematen zaien bilatu genuen, eta, irizpide linguistiko eta estatistikoak uztartuta, ordain bana hautatu genuen UF bakoitzeko. Asmatze-tasa on samarra lortu genuen patroi morfosintaktikoak eta UFen ordainak automatikoki esleitzerakoan, eta ikusi genuen datu automatiko horiek ere oso lagungarriak zirela bai identifikazio- eta bai itzulpen-lanerako, eskuz aztertutakoak bezain lagungarriak ia.

Hortaz, prest geneukan eskalabilitate-arazoa konpontzeko proposamena ere. Gorabeherak gorabehera, lan dezente egina genuen dagoeneko, eta gero eta gertuago nuen amaiera.

5. Conference of the European chapter of the Association for Computational Linguistics

6. *Corpus anotatu* esaten zaio hizkuntza-anotazioak dauzkan corpusari. Anotazioak askotarikoak izan daitezke (adibidez, hitz bakoitzaren kategoria gramatikala zein den, hitzen arteko erlazio sintaktikoa zein den…), baina, kasu honetan, UFak dira anotatutakoak.

## Nahaspilak forma hartu zuenekoa

Uste dut tesigintzaren urratsak ez direla berdinak izaten alor guztietan, baina, hizkuntzalaritza konputazionaleko tesi gehienetan, beste alor tekniko askotan bezala, lehenik ikerketa-lana egiten da, eta bukaera-bukaerarako uzten da tesiaren idazketa. Nik argitaratuak nituen artikulu batzuk han eta hemen, baina eduki hura guztia batzea, ordenatzea eta nahaspilari forma ematea falta zen. Eta, aizue, horretan hasi orduko, erabat bestelakoa iruditu zitzaidan dena.

Izan ere, tesiaren motibazio nagusia itzulpen automatikoaren bueltan ezarri genuen hasieran, eta, lehen ere esan dudanez, sistema neuronalen sorrera tarteko, alor horretan eginiko ekarpena oso txikia iruditzen zitzaidan. Ez nuen gogoan hartzen, ordea, ildo nagusi horrez gain beste ildo paralelo bat ere ondu genuela aldi berean, baliabideei zegokiena.

Hasteko, datu-base bat sortu genuen eginiko azterketan bildutako informazio guztia gordetzeko, *Konbitzul*,[7] eta edonoren eskura jarri genuen, Interneten, kontsultak egiteko interfaze eta guzti. LREC[8] kongresuan aurkeztu genuen, eta kongresuko artikulu-bilduman argitaratu genituen baliabide horri buruzko xehetasunak (Iñurrieta *et al.*, 2018b). Gaztelaniazko 1.927 UF eta euskarazko 2.074 jasotzen ditu *Konbitzul*ek, gutxienez ordain banarekin batera, eta honako kontsulta hauek egiteko aukera ematen du, adibidez:

- Gaztelaniazko aditza+izena motako UF bati zer ordain ematen zaion euskaraz (edo euskarazko bati gaztelaniaz).
- Izen jakin batek zer aditzekin osatzen dituen UFak (edo aditz jakin batek zer izenekin).
- Gaztelaniazko aditza+izena motako UFrik usuenek zer ezaugarri morfosintaktiko dituzten.
- Gaztelaniazko UF horien euskarazko ordainek zer ezaugarri lexiko eta morfosintaktiko dituzten.

Horrez gain, aukera dago xeheki aztertutako 894na UF eta ordainen inguruko informazioa taulatan deskargatzeko, ikerketan edo hizkuntza-tresnen garapenean erabili nahi dutenentzat.

Bestalde, hizkuntza-tresnak sortzeko eta ebaluatzeko, corpusek berebiziko garrantzia dute, eta bi corpus anotatu osatzen lagundu genuen guk, artikulu honen hasiera samarrean aipatu dudan PARSEME proiekturako. Proiektuaren helburuen artean zegoen fraseologia konputazionaleko ikerketa bultzatzeko baliabideak sortzea, eta asmo horrek bultzatuta egin zuten proposamena. Hiru fasetan landu da corpusa, eta hiru argitaraldi izan ditu: lehenengoan, gaztelaniazko corpusaren anotatzaile-taldean parte hartu genuen, hau da, testuak aztertu eta UFak banan-banan markatu eta sailkatu genituen gaztelaniazko testuetan (Savary *et al.,* 2018); bigarrenean, berriz, euskarazko corpusa ere sortzea erabaki genuen, eta, Ixako hizkuntzalari-taldetxo

7. http://ixa2.si.ehu.es/konbitzul

8. Language Resources and Evaluation Conference

bat elkartuta, 3.823 UF anotatu genituen guztira 11.158 esaldiko corpusean; hirugarren fasea oraintsu amaitu berri dugu, eta aurreko corpusen bertsio findua da.

Euskarazko corpusa sortzea interesgarria izan zen, baliabide-sorkuntzaren ikuspuntutik ez ezik, hizkuntzalaritzaren ikuspuntutik ere, hogei hizkuntzatako corpusak anotatu baikenituen gidalerro berberei jarraituz (Ramisch *et al.*, 2018). Horrek aukera eman zigun, besteak beste, agerian jartzeko euskarazko UFek beste hizkuntza askoren aldean dituzten zenbait ezaugarri bereizgarri. Deigarriena aditz arindun konbinazioen maiztasuna da, ziur asko (*lan egin, lo hartu* eta halakoena): euskaraz, 100 esalditik 34k dute halakoren bat, batez beste; frantsesez eta gaztelaniaz, 100etik 20k eta 15ek, hurrenez hurren; eta ingelesez, 100etik 6k. Corpuseko hizkuntza guztiak kontuan hartuta, batezbestekoa 100 esalditik 18koa da, eta bi hizkuntzak baino ez dute euskarak baino maiztasun altuagoa: hindiak eta persierak. Artikulu batean jaso genituen era horretako datuak eta beste batzuk, gidalerroak euskarara egokitzeko zailtasunak (eta ezintasunak) barne, eta fraseologia konputazionalari buruzko lantegi batean azaldu genituen (Iñurrieta *et al.,* 2018c).

Gainera, euskarazko corpusa sortu izanak beste ate bat ireki zidan lankidetzarako: PARSEMEn arduradun-lanetan ibilitako bi ikerlarik azterketa bat egin nahi zuten familia filogenetiko desberdineko zenbait hizkuntzatan, eta euskara ere aztergaien artean sartu nahi zuten. Amuari heldu, eta bost hizkuntzatan egin genuen azterketa: alemanez, euskaraz, grezieraz, polonieraz eta portugesez. UFak anotatuta zeuzkaten corpusetatik abiatuta, agerpen literalei erreparatu genien, hau da, UFaren osagai-hitzak literalki erabiltzen diren kasuei.

Honako hau zen abiapuntuko hipotesia: hitz-konbinazio jakin bat UFa izan badaiteke, hitz-konbinazio hori oso gutxitan erabiltzen da literalki, eta, hala erabiltzen denean, gehiengehienetan, ezaugarri morfosintaktikoei begiratuta bereiz daiteke agerpena idiomatikoa ala literala den. Esate baterako, *ziri* eta *sartu* hitzak idiomatikoki erabilita daude 11. adibidean, baina literalki 12.ean:

(11) Ez zen benetan ari. *Ziria sartu* zizun!
(12) Mutikoak egurrezko *ziri* bat *sartu* zuen zuloan.

Lan horren bidez erakutsi genuen bigarren erabileraren gisakoak oso urriak direla eta, gehienetan, izaten dela zantzu morfosintaktikoren bat agerpena literala dela automatikoki jakiteko; kasu honetan, izena *-a* artikulu mugatuarekin agertu beharrean *bat* determinatzailearekin agertzea. Hain dira urriak halakoak, ezen, corpusean UF gisa markatutako hitz-konbinazioak oinarritzat harturik, UF horietako osagai-hitzak elkarrekin agertzen diren kasu guztien % 2 bakarrik baitira literalak. Gainera, morfosintaxiari halako garrantzia eman genionez, lan hori erdiz erdi zetorkidan tesian sartzeko, itzultzaile automatikoetarako eginiko proposamenari indarra ematen baitzion, nolabait. *Prague Bulletin of Mathematical Linguistics* aldizkarian argitaratu genuen haren berri (Savary *et al.*, 2019a), eta euskarazko atalaren bertsio egokitu bat aurkeztu nuen IkerGazten (Iñurrieta, 2019a).

Hala, hasierako ildo nagusitik kanpoko lan horiekin osatu nituen tesiko bi kapitulu, eta, gauzak zer diren, hasieran garrantzirik gabekoak iruditzen zitzaizkidan lantxo horiek izan dira, azkenean, nire atalik gustukoenak.

### Hautsi-urratuak egin genituenekoa

Idatzi eta zuzendu, idatzi eta zuzendu, buelta batzuk eman ondoren, erabaki nuen bazela garaia lana bukatutzat emateko. Ez, noski, txostena erabat biribilduta zegoela pentsatzen nuelako (ez dakit baden tesigilerik sentipen horrekin amaitzen duenik), baizik eta idazketa-zuzenketa prozesuak luzeegi joko lukeelako mugarik jarri ezean eta, batez ere, bukatzeko gogo itzela nuelako. Joan den urteko irailean utzi nituen bi txostenak gordailuan: euskarazkoa, tesi osoa, eta ingelesezkoa, tesiaren laburpen batekin eta ingelesez eginiko argitalpenekin osatua.

- *Aditza+izena Unitate Fraseologikoak gaztelaniatik euskarara: azterketa eta tratamendu konputazionoala* (Iñurrieta, 2019b)
- *Verb+Noun Multiword Expressions: a linguistic analysis for identification and translation* (Iñurrieta, 2019c)

Defentsa, berriz, handik bi hilabetera izan zen, Informatika Fakultatean, Rikardo Etxeparek (Centre National de la Recherche Scientifique), Margarita Alonsok (Universidade da Coruña) eta Miren Azkaratek (Euskal Herriko Unibertsitatea) osatutako epaimahaiaren aurrean. Nerbioak dantzan igaro nituen aurkezpena hasi aurreko orduak, baina, bitxia bada ere, behin aretoan sartutakoan, urduritasunak alde egin zuen, eta oso atsegina izan zen dena.

Badakidan arren argi-itzalez beteriko bidea izan dela, eta nahi baino luzeagoa, zorionez, asmatzen dugu, batzuetan, gaizki eginak geure buruari barkatzen eta ongi eginak kontuan hartzen. Gaur egun, atzera begiratzen dudanean, argi gehiago ikusten ditut eginiko bidean, eta itzalak ere ez zaizkit lehen bezain ilunak iruditzen. Azken batean, doktoretza-tesi bat egitea ikasketa-prozesu bat baita, eta hura bukatutakoan hasten omen da ikertzaile baten benetako ibilbidea. Hala izan bedi!

## ERREFERENTZIAK


Etchegoyhen, Thierry; Eva Martínez; Andoni Azpeitia, Gorka Labaka & Iñaki Alegria (2018). «Neural machine translation of Basque», *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alicante, Espainia, 139.-148. or.

Gurrutxaga, Antton (2014). *Idiomatikotasunaren katakterizazio automatikoa: izena+aditza konbinazioak*. Doktoretza-tesia, UPV/EHU.

Iñurrieta, Uxoa (2019a). «Unitate fraseologikoen agerpen literalak, urre baina urri», *IkerGazte: nazioarteko ikerketa euskaraz. Kongresuko artikulu-bilduma. Giza zientziak eta artea*, Baiona, 139.-147. or.

——— (2019b). *Aditza+izena Unitate Fraseologikoak gaztelaniatik euskarara: azterketa eta tratamendu konputazionala*. Doktoretza-tesia, UPV/EHU.

——— (2019c). *Verb+Noun Multiword Expressions: a linguistic analysis for identification and translation.* PhD thesis, UPV/EHU.

Iñurrieta, Uxoa; Itziar Aduriz; Arantza Díaz de Ilarraza; Gorka Labaka & Kepa Sarasola (2014). «Izen+aditz konbinazioen azterketa elebiduna, hizkuntza-aplikazio aurreratuei begira», *Linguamática 6(2)*, 45.-55. or.

——— (2016a) «Izen+aditz konbinazioen itzulpenaz eta tratamendu konputazionalaz», *Senez* 47, 237.-249. or.

——— (2017) «Rule-based translation of Spanish verb-noun combinations into Basque», *Proceedings of the 13th Workshop on Multiword Expressions (at EACL 2017)*, Valentzia, 149.-154. or.

——— (2018a) «Analysing linguistic information about word combinations for a Spanish-Basque rule-based Machine Translation system», *Multiword Units in Machine Translation and Translation Technologies*, John Benjamins Publishing Company, 41.-60. or.

——— (2018b). «Konbitzul: an MWE-specific database for Spanish-Basque», *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japonia, 2.500.-2.504. or.

Iñurrieta, Uxoa; Itziar Aduriz; Arantza Díaz de Ilarraza; Gorka Labaka, Kepa Sarasola & John A. Carroll (2016b). «Using linguistic data for English and Spanish verb-noun-combination identification», *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016): Technical papers*, Osaka, Japonia, 857.-867. or.

Iñurrieta, Uxoa; Itziar Aduriz; Ainara Estarrona, Itziar Gonzalez-Dios; Antton Gurrutxaga; Ruben Urizar & Iñaki Alegria (2018c). «Verbal multiword expressions in

Basque corpora», *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (at COLING 2018)*, Santa Fe, AEB, 86.-95. or.

Mayor, Aingeru; Iñaki Alegria; Arantza Díaz de Ilarraza; Gorka Labaka; Mikel Lersundi & Kepa Sarasola (2009). «Matxin, euskararako lehenengo itzultzaile automatikoa». *Senez* 37, 197.-220. or.

Ramisch, Carlos; Silvio R. Cordeiro; Agata Savary; Veronika Vincze; Verginica B. Mititelu *et al.* (2018). «Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions». *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (at COLING 2018)*, Santa Fe, AEB, 222.-240. or.

Savary, Agata; Marie Candito; Verginica B. Mititelu; Eduard Bejcek; Fabienne Cap *et al.* (2018). «PARSEME multilingual corpus of Verbal Multiword Expressions». *Multiword Expressions at length and in-depth: extended papers from the MWE 2017 workshop*, Language Science Press, 87.-147. or.

Savary, Agata; Silvio R. Cordeiro; Timm Lichte; Carlos Ramisch; Uxoa Iñurrieta & Voula Giouli (2019). «Literal occurrences of multiword expressions: rare birds that cause a stir». *Prague Bulletin of Mathematical Linguistics* 112, 1.-44. or.

Savary, Agata; Mamfred Sailer; Yanick Parmentier; Michael Rosner; Veronica Rosén *et al.* (2015). «PARSEME—PARSing and Multiword Expressions within a European multilingual network». *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*. Poznan, Polonia.

Urizar, Ruben (2012). *Euskal lokuzioen tratamendu konputazionala*. Doktoretza-tesia, Informatika Fakultatea, UPV/EHU.

**Resumen**

Este es un artículo no académico sobre un trabajo académico. Recién concluida su tesis doctoral sobre lingüística computacional, Uxoa Iñurrieta da cuenta en estas páginas del camino recorrido desde el inicio del doctorado hasta su finalización. Sin embargo, el hilo conductor del artículo, más que el trabajo de investigación en sí, lo constituye la trayectoria de la propia doctoranda, quien se vale de esta narración para dar a conocer las ideas principales de su tesis.

**Résumé**

Ceci est un article non-académique sur un travail académique. Uxoa Iñurrieta a récemment défendu sa thèse de doctorat sur la linguistique computationnelle et retrace ici le parcours de son doctorat, mais plutôt que de placer son travail de recherche au centre de son propos, elle nous présente ici ces aléas de chercheuse, et, à travers ce récit, expose les idées principales de sa thèse.

**Abstract**

This is a non-academic article about an academic work. Having recently completed her doctoral dissertation on computational linguistics, Uxoa Iñurrieta recounts her journey from the beginning to the end of her doctoral studies. Nevertheless, the guiding thread of the article is not the doctoral candidate's research, but rather her academic career, though this narration does include the main ideas of her dissertation.