Izen + aditz konbinazioen itzulpenaz eta tratamendu konputazionalaz

Uxoa Iñurrieta, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola

Artikulu hau irakurriko duen orok badaki —esperientziaz, ziur asko— testu bat itzultzeko, ez dela nahikoa testuko hitzak eta haien ordainak ezagutzea. Esaldi gutxi batzuk hitzez hitz itzul daitezke, baina, gehienetan, hizkuntza batetik bestera hitzez hitz ekarritakoak oso traketsak izaten dira:

(1) Hizlariak berehala *gatibatu* zuen entzuleen interesa → interesa *bereganatu* zuen

(2) Epeak datorren astean *garaituko du* → *epea amaituko da*

Unitate Fraseologikoak (UFak) hizkuntzatik hizkuntzara asko aldatzen diren itzulgaiak dira, eta, hain zuzen ere, horiexek sortzen dute arazoa 1. eta 2. adibideetan. Bi esaldiak dira gaztelaniatik automatikoki itzuliak, eta, ziur asko, nahiko xelebreak irudituko zaizkio euskara pixka bat dakien edonori.

Erdaratik hitzez hitz kalkatutako beste batzuk, ordea, ezagunagoak egiten zaizkigu, nahiko zabalduta baitaude hainbat euskaldunen artean. Ibon Sarasolak (1997) bildutako *aje*etako batzuek, adibidez, zerikusia dute UFekin:

(3) atentzioa deitu/prestatu → atentzioa eman

(4) ilea hartu → adarra jo

(5) kontutan izan/hartu/eduki → kontuan izan/hartu/eduki

Hiztun batzuei zaila egiten zaie halakoak ondo erabiltzea, erdaraz normal-normalak diren esapideak euskarara bere horretan ekartzen saiatzen direlako, eta, beharbada, ez dakitelako euskaraz badaudela forma jatorragoak gauza bera adierazteko. Bada, aplikazio informatikoei ere antzeko zerbait gertatzen zaie, eta horretaz arituko gara gu lan honetan. Izen batez eta aditz batez osatutako konbinazioak izango ditugu aztergai, 1-5 adibideetakoak bezalakoak.

Hasteko, fraseologiaren inguruko oinarrizko kontzeptu batzuk argitzen saiatuko gara, eta, gero, gure lana zertan datzan azalduko dugu: UFek itzultzaile automatikoei zer-nolako zailtasunak sortzen dizkieten, zailtasun horiei nola egin nahi diegun aurre, eta sortu berri dugun baliabidea nolakoa den.

Unitate Fraseologikoen inguruko hainbat kontzeptu oinarrizko

XVI. mendean *Refranes y sentencias* idatzi zenetik, esaera- eta atsotitz-bilduma bat baino gehiago argitaratu da euskaraz: Mokoroaren *Ortik eta Emendik* (1990)[1], Intza Proiektuaren *Euskal Lokuzioak Sarean*[2], eta Garateren *Atsotitzak*[3], sarean eskuragarri dauden batzuk aipatzearren. Bildumez gain, ordea, ezin da esan gurean lan asko egin denik fraseologiaren inguruan.

Gaur egun, badirudi gero eta ikerketa gehiago egiten ari dela alor horretan, bai beste hizkuntza batzuetan, bai euskaraz, eta, ikerketa-lanak ugaritzearekin batera, UFak sailkatzeko irizpideak ere findu egin dira. Azken lau urteotan, besteak beste, euskarazko fraseologiaren inguruko hiru doktoretza-tesi argitaratu dira:

- *Euskal lokuzioen tratamendu konputazionala*, Ruben Urizar (2012).
- *Unitate fraseologikoen itzulpena: alemana-euskara. Literatur testuen corpusean oinarritutako analisia,* Zuriñe Sanz (2014).
- *Idiomatikotasunaren karakterizazio automatikoa: izena + aditza konbinazioak,* Antton Gurrutxaga (2015).

Esan genezake gure lanak hirurekin duela, nolabait, zerikusia: batetik, aplikazio informatikoei begira ari garelako; bestetik, erdaratik euskararako itzulpenean oinarritzen dugulako gure azterketa; eta, hirugarrenik, gure ikergai nagusia izen + aditz motako konbinazioak direlako.

Urizar eta Gurrutxaga bezalaxe, gu ere Corpasen (1997) sailkapenean oinarrituko gara terminologia-irizpideei dagokienez, eta lokuzioak eta kolokazioak bereiziko ditugu[4]. Unitate Fraseologikotzat hartzen dira bi konbinazio motak, hau da: hitz batez baino gehiagoz osatutako esanahi-unitateak dira, eta idiomatikoak dira nolabait, idiomatikotasun hori maila batekoa edo bestekoa izan daitekeen arren (Sag *et al.*, 2002).

Azken ezaugarri horri begiratuz gero, **lokuzioak** dira idiomatikoenak, *adarra jo* bezalakoak; esanahi aldetik opakoak dira, eta, azpian dituzten hitzei begiratuta bakarrik, ezin da jakin konbinazio osoak zer esan nahi duen. **Kolokazioetan**, berriz, hitzetako batek bere esanahia gorde-

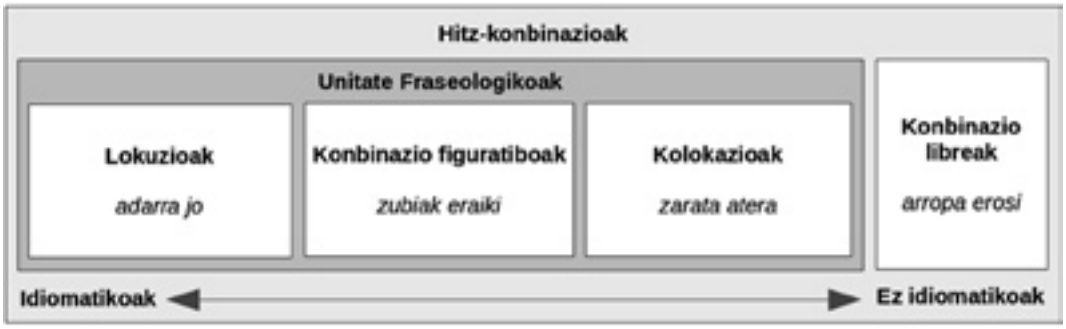
1. http://www.erabili.eus/lantresnak/esamoldeak/mokoroa
2. http://intza.armiarma.eus/
3. http://www.ametza.com/bbk/htdocs/garate.htm
4. Enuntziatu fraseologikoak (atsotitzak, aipuak, errutinazko formulak…) alde batera utziko ditugu, gure ikergai diren izen + aditz konbinazioak ez baitira multzo horretakoak izaten. Dena den, halakoen inguruko informazio gehiago nahi izanez gero, jo Urizarren tesiko 84.-89. orrialdeetara.

tzen du normalean, eta besteak esanahi horren lagungarri gisa jokatzen du, *atentzioa eman* edo *zarata atera*rekin gertatzen den bezala. Kolokazioen parte diren hitzek ohi baino joera handiagoa izaten dute elkarrekin agertzeko, eta hizkuntza bat jariotasunez hitz egiten duenak arazorik gabe erabili ohi ditu (Gurrutxaga, 2015).

Izena + *egin/eman/hartu* konbinazioak ere, *lan egin* edo *min hartu* bezalakoak, kolokazioen multzoan sartzen dira gehienetan, baina badute beste berezitasun bat ere: horrelakoetan agertzen diren aditzak *arinak* direla esaten da (Zabala, 2004), esanahia *arindu* egiten zaielako; izenari aditz izaera emateko balio dute nolabait, eta haren esanahiari ñabarduraren bat gehitzeko, baina berena guztiz gorde gabe.

Bereizketa horrez gain, **konbinazio figuratiboak** ere beste multzo batean sartzen ditugu guk, Howarthi (1996) jarraituz, besteak beste. Idiomatikotasunaren continuumean, lokuzioen eta kolokazioen arteko punturen batean leudeke hirugarren multzo horretakoak, lehenengoetatik gertuxeago. Izan ere, lokuzioetan bezala, konbinazio figuratiboen esanahi osoa ez da beti azpiko hitzen banakako esanahien batura izaten, baina metaforikoki uler litezke askotan, *zubiak eraiki*rekin gertatzen den moduan, non zubia fisikoki edo metaforikoki eraiki daitekeen —adibidez, bi herrialderen arteko harremanei buruz hitz egitean—.

Azkenik, Unitate Fraseologikoetatik kanpo geratzen diren hitz-konbinazioak **libreak** direla esaten da, hitz bakoitzak bere esanahia gordetzen baitu, eta konbinazioek ez baitute inongo berezitasunik idiomatikotasunari dagokionez: *mendira joan, oparia erosi, kalean ibili*…



1. irudia: Hitz-konbinazioen sailkapena.

Izen + aditz konbinazioak gaztelaniatik euskarara eta euskaratik gaztelaniara

Esan dugunez, UFak asko aldatzen dira hizkuntza batetik bestera, eta, jakina, hizkuntzak zenbat eta tipologikoki urrunago egon elkarrengandik, desberdintasunak orduan eta nabarmenagoak dira. Guk, beste ezeri ekin aurretik, gaztelaniazko eta euskarazko UFen arteko desberdintasun horien irudi orokor bat lortu nahi izan genuen, eta hiztegi elebidunetara jo genuen horretarako (Iñurrieta *et al.*, 2014).

Izen batez eta aditz batez osatutako konbinazioak aukeratu genituen gure azterketarako, bi arrazoigatik: batetik, oso sarri erabiltzen direlako, eta, bestetik, sintaktikoki nahiko malguak izan ohi direlako eta horrek bereziki zailtzen duelako haien tratamendu konputazionala, geroago azalduko dugun bezala. Elhuyarren gaztelania-euskara hiztegia[5] hartu, eta hizkuntza bateko zein besteko konbinazio-zerrendak erauzi genituen, ordainekin batera.

Euskarazko konbinazioen artean, denak ziren izen batez eta aditz batez osatuak, baina izenek kasu- edo postposizio-marka bat baino gehiago zituzten, hiru laurden baino gehiago absolutiboan bazeuden ere: *arnasa hartu, eguzkiak erre, bistan izan, gehiegikeriaz jokatu…* Gaztelaniazko konbinazioetan, berriz, preposizioak eta determinatzaileak ere onartu genituen bi osagai nagusien artean, euskarazko izenei lotutako marken parekoak zirelakoan: *causar asombro, cerrar la noche, decir entre dientes, estar al quite…*

Espero bezala, azterketatik ateratako datuek argi erakusten dute izen + aditz konbinazioen ordainak, euskaratik gaztelaniara eta gaztelaniatik euskarara, oso gutxitan direla barruko osagaien banakako ordainez osatuak, hau da, konbinazio osoa itzultzeko, gehienetan, ez zaizkiela izenari, aditzari eta konbinazioa osatzen duten beste morfemei zeini bere ordaina ematen. Hona hemen datu estatistiko batzuk:

- Gaztelaniazko konbinazioen artean, %48,54k bakarrik dute ordaintzat euskarazko izen + aditz motako konbinazio bat. Hortaz, gehienetan, euskarazko ordainaren egitura morfologikoa ez da gaztelaniazkoaren parekoa (6), eta, hala denean ere, %78,21 kasutan, bi hitzetako batek behintzat ez du bere ohiko ordaina gordetzen itzulitakoan (7):
  (6) alzar *el vuelo* → *hegan* hasi (adberbioa + aditza)
  (7) *vencer* el plazo → epea *amaitu* (eta ez *garaitu*)
- Euskarazko konbinazioen ordainei begiratuta, are nabarmenagoa da aldea: %30,85 bakarrik dira aditz + (preposizio) + (determinatzaile) + izen motakoak, eta oso deigarria da ordain guztien %58,07 aditz soilak direla, eta ez hitz-konbinazioak (8). Bestalde, morfologia aldetik parekoak diren horien artean ere, %28,01 kasutan bakarrik dira hizkuntza bateko eta besteko aditzak eta izenak ordainak (9):
  (8) negar egin → llorar (aditza bakarrik)
  (9) barrez *ito* → *morirse* de risa (eta ez *ahogarse*)

Beraz, hori guztia kontuan hartuta, esan genezake aditzez eta izenez osatutako UFak oso gutxitan itzultzen direla hitzez hitz gaztelaniaren eta euskararen artean. Zehazki, gaztelaniatik euskarara, %10,57 izan dira hitzez hitzezko ordainak, eta, euskaratik gaztelaniara, berriz, %8,64.

Noski, hiztegietan begiratu beharrean corpusetan bilatuko bagenu, zenbakiak aldatu egingo lirateke ziur asko, kontuan izan behar baita hiztegiek gordetzen dituzten konbinazioen kopurua

5. hiztegiak.elhuyar.eus

mugatua dela, eta, gainera, lokuzioak jasotzen dituztela batez ere —hain zuzen ere, hizkuntza batetik bestera gehien aldatzen diren UFak—. Dena den, azterketatik ateratako zenbaki horiek UFak itzultzearen konplexutasunaren zantzu bat ematen digute, eta, beharbada, lagungarriak izango dira ulertzeko aplikazio informatikoek zergatik sortzen dituzten hainbeste akats halakoak prozesatzean.

UFen itzulpen automatikoa: hutsuneak eta erronkak

UFek alor batean baino gehiagotan sortzen dituzte arazoak: atzerriko hizkuntzen ikasketan, hizkuntzalaritza orokorrean, hiztegigintzan… Gu, hemen, hizkuntzaren tratamendu konputazionalaz arituko gara; itzulpen automatikoaz (IA), zehazki.

Gure lanerako erabiltzen ari garen itzultzaile automatikoa Matxin da, arau linguistikoetan oinarritutako sistema bat (Mayor *et al*., 2009). Labur-labur, honako teknika hau darabil Matxinek testuak itzultzeko:

a) Gaztelaniazko testua jaso, eta linguistikoki analizatzen du: kategoria gramatikalak, funtzio sintaktikoak, hitzen barruko morfemak…
b) Hiztegi elebidun batez baliatuta, lema eta morfema bakoitzari euskarazko ordainak ematen dizkio.
c) Euskarazko lemetatik hitzak sortu, sintagmak osatu, eta esaldia egituratzen du.

Matxinen oinarrian dagoen hiztegi elebidunak baditu hitz-konbinazioak diren sarrera batzuk, baina ez dira asko, eta, gainera, oso modu sinplean prozesatzen dira oraindik. Batetik, gaztelaniazko itzulgaia analizatzean, begiratzen da ea esaldian badagoen hiztegiko hitz-konbinaziorik, baina ez dira beti ondo detektatzen: hitz guztiak beti jarraian bilatzen dira, eta hurrenkera eta forma berean —aditzaren flexioa bakarrik hartzen da kontuan—. Horrek esan nahi du, adibidez, hitz-konbinazioko osagaien artean beste hitzen bat sartuz gero (10b), ez dela hitz-konbinazioa UFtzat hartzen eta hitz bakoitza bere aldetik itzultzen dela:

(10a) ES: Las ingenieras *llevaron a cabo* el proyecto.
IA: Ingeniariek proiektua *burutu* zuten.

(10b) ES: Las ingenieras *llevaron* el proyecto *a cabo*.
IA: *Ingeniariek proiektua *eraman zuten kabora*.

Bestetik, behin gaztelaniazko UFa detektatutakoan, beti euskarazko ordain berbera ematen zaio konbinazioari (11b), eta ez zaio azpiko hitzen informazio linguistikoari begiratzen (11c). Horrek ere, noski, akatsak sorrarazten dizkio sistemari:

(11a) ES: El profesor *se da importancia* en clase.
IA: Irakaslea klasean *handiustea da*.

(11b) ES: En clase, siempre se da importancia a lo mismo.

IA: *Klasean, beti betikoa handiustea da.

(11c) ES: Los profesores *se dan importancia* en clase.

IA: *Irakasleak klasean *handiustea da*.

Hortaz, bi erronkari egin behar zaie aurre: batetik, sorburu-hizkuntzako UFen detekzioa hobetzeari, eta, bestetik, detektatutako UF horiek euskarara ekartzeko modu eraginkorragoak bilatzeari. Adibideek agerian uzten dutenez, informazio linguistiko gehigarria behar-beharrezkoa da ataza baterako zein besterako.

1. Informazio sintaktikoa, gaztelaniazko UFak automatikoki detektatzeko baliagarri

UFak automatikoki detektatzea ez da batere lan erraza, oreka bilatu behar baita arau zorrotzegien eta lausoegien artean. Matxinek orain erabiltzen duen metodoak oso ongi detektatzen ditu UF batzuk, baina kanpoan uzten ditu beste asko. Demagun gaztelaniazko bi esaldi hauek ditugula:

(12a) No te entiendo, *haz un esfuerzo* por explicarte.

(12b) Ni todo *el esfuerzo que hizo* le sirvió para hacerse entender.

Matxinen hiztegi elebidunean badago *hacer un esfuerzo* sarrera, eta, horri esker, lehen adibidea (12a) arazorik gabe detektatzen da, hitz-konbinazioa hiztegian bezalaxe agertzen baita esaldian, aditzaren flexioa bakarrik aldatuta. Bigarrenean (12b), ordea, gaur egungo detekzio-sistema ez da gai UFrik ezagutzeko, *hacer un esfuerzo* konbinazioa ez delako bere horretan agertzen: izena eta aditza alderantzizko hurrenkeran daude, eta izen-sintagma zehaztua da —*el* determinatzailea darama, *un* beharrean—. Gaur egungo metodoa itxiegia da, hortaz.

Bestalde, metodo zabalegiak erabiltzea ere ez da komeni, nahi baino hitz-konbinazio gehiago detektatzeak ere akatsak sorraraziko bailizkioke sistemari. Esate baterako, izenaren eta aditzaren lemak bakarrik bilatuko balitu, besterik gabe, honelako esaldiak zeharo nahasgarriak izan litezke Matxinentzat:

(12c) Le supuso *un esfuerzo hacerse* entender.

Suponer un esfuerzo ezagutu beharrean, *hacer un esfuerzo* detektatuko luke beharbada, edo bai bata eta bai bestea, eta esaldi guztia gaizki analizatuko luke horren ondorioz. Beraz, bistan da beste moduren bat bilatu behar dela UFak ongi prozesatu ahal izateko.

Hori guztia kontuan izanik, gaztelaniazko konbinazio sorta bat hartu, eta informazio sintaktikoa aztertzeari ekin genion, informazio hori Matxini gehitzea lagungarria izan ote zitekeen jakiteko. Hiztegitik hartutako konbinazioak corpus paralelo handi batean bilatu, eta sarrien errepikatzen ziren 150ak hautatu genituen.

Lehenik eta behin, lexiko-semantikaren ikuspuntutik sailkatu genituen, iruditu baitzitzaigun multzo horretako batzuek ez zutela benetan tratamendu berezirik behar. Lokuzioak, kon-

binazio figuratiboak, kolokazioak eta konbinazio libreak bereizi genituen, eta azken multzokoak alde batera utzi. Guztira, 99 geratu zitzaizkigun.

Konbinazio bakoitzaren hainbat ezaugarriri begiratu genien: preposizioei, determinatzaileei, izen-sintagmaren zehaztasunari eta numeroari, funtzio sintaktikoari, aditza eta izen-sintagma bereizteko aukerari, eta hitz-hurrenkera aldatzeko aukerari. Taulatan bildu genituen datu horiek guztiak:

	Prep	**Det**	**Zeh**	**Num**	**Sint**	**Bereiz**	**Hurr**
Hacer un esfuerzo	-	aukeran	aukeran	aukeran	obj	bai	bai
Estar en vigor	en	ez	mg	s	mod	bai	ez

1. taula: Informazio sintaktikoa jasotzeko taulen adibide bat.

Jasotako datuen arabera, *hacer un esfuerzo* konbinazioa sintaktikoki librea da: determinatzailea eraman dezake, baina ez du derrigorrezkoa; zehaztasuna eta numeroa aukerakoak ditu; izen-sintagma eta aditza bereiz daitezke, eta hitz-hurrenkera ere aldakorra du. *Estar en vigor*, aldiz, ez da hain konbinazio malgua, guztiz finkoa ere ez den arren: izen-sintagmak ezin du determinatzailerik izan, eta beti doa singularrean; izen-sintagma eta aditza bereiz daitezke, baina ezin da hitz-hurrenkera aldatu.

Hizkuntzalaritza konputazionalean, bi neurri erabili ohi dira sistemen emaitzak aztertzeko: doitasuna (*precision*) eta estaldura (*recall*). Doitasuna deitzen diogu, gure kasuan, detektatutako hitz-konbinazioen artean zuzen zenbat detektatu diren adierazten duen neurriari; eta estaldura, berriz, detektatu behar ziren UF guztietatik zenbat detektatu diren adierazten duenari. Matxinek gaur egun UFak detektatzeko darabilen metodoak oso doitasun altua du, baina estaldura txikia. Esperimentu batzuk egin ondoren, ikusi dugu, gure lanean aztertutako datuak erabilita, estaldura asko igotzen dela (% 30 inguru), eta doitasuna oso gutxi jaitsi (% 1,5 inguru). Beraz, baiezta dezakegu informazio linguistikoa, oro har, oso lagungarria dela UFen detekzioa hobetzeko.

2. UFak euskarara ekartzea: hautapen lexikala eta gramatika

Zeregina, noski, ez da hor amaitzen, hitz-konbinazioak ezagutu eta gero euskarara itzuli behar baitira, eta hori ere ez da batere lan erraza Matxin eta halako itzultzaile automatikoentzat. Kontuan izan behar da, detekzioan bezala, euskaratze-prozesuan ere UFak hitz bakarra balira bezala tratatzen direla Matxinen, azpian zer duten begiratu gabe.

(13a) ES: Es fácil *perder los estribos* en esa situación.
IA: Erraza da egoera horretan *nor bere onetik ateratzea*.

(13b) ES: Ella *perdió los estribos*.
IA: **Hark nor bere onetik atera zuen*.

Bi itzulpen horietan, lehena zuzena da, Matxinen hiztegi elebidunean *perder los estribos = nor bere onetik atera* agertzen baita. Baina sistemak itsu-itsuan erabiltzen du hiztegiko informazioa; ez daki UFen azpian dauden hitzek zer kategoria duten eta nola erabili behar diren, eta horrek bigarren adibidekoa (13b) bezalako itzulpenak sortzea eragiten dio.

Hortaz, euskaratze-prozesura begira ere, garrantzitsua da informazio linguistikoa kontuan hartzea. Batzuetan, akatsa lexikoa hautatzean egiten da:

(13) ES: Ahora todo está *en juego*.
IA: Orain dena *jolasean* dago.

Horrelakoetan, UFaren barruko hitzen ordain egokiak zein diren esan behar zaio sistemari: *estar en juego* aurkituz gero, erabili *joko* izena, eta ez *jolas*. Beste batzuetan, berriz, UFen ondorioz sortzen diren akatsak gramatikalak dira, eta beste informazio mota bat da beharrezkoa:

(14) ES: Han cambiado *de tema*.
IA: *Gaitik* aldatu *dira.*

Matxinek jakin behar duena da: *cambiar de tema* euskaratzean, absolutiboa erabiltzen dela ablatiboaren ordez —*gaia* aldatu, eta ez *gaitik*—, eta izena aditzaren objektu bihurtzen dela —beraz, gaia aldatu *dute*, eta ez *dira*—.

Bestalde, aurreko atalean azaldu dugunez, kontuan izan behar da gaztelaniazko UFen ordainak ez direla beti beste UF batzuk izaten euskaraz, eta, horrelakoetan ere, informazioa egokitu beharra dago:

(15) ES: La pareja *contrajo matrimonio*.
IA: Bikotea *ezkontza uzkurtu* zen.

Azken adibide horretan, sistemak jakin beharrekoa ez da lexiko mailakoa edo gramatika mailakoa bakarrik, baizik eta biei buruzkoa: batetik, izenak ez du batere ordainik behar, konbinazio osoaren ordaina aditz bakarra baita; eta, bestetik, aditza euskarara ekartzean ez da *uzkurtu* erabili behar, baizik eta *ezkondu*.

Lehenik eta behin, behar duten tratamenduaren arabera sailkatu ditugu detekziorako landuak genituen 99 UFak: informazio lexikala bakarrik behar dutenak, gramatikala bakarrik behar dutenak, eta bai bata eta bai bestea behar dituztenak. Eta, ondoren, kasu bakoitzean gehitu behar zaizkien datuak aztertu ditugu, 13.-15. adibideetarako aipatu ditugunak bezalakoak.

	ize	**adi**	**IS kas/pos**	**Num**	**Mug**	**Sint**
Cambiar de tema	-	-	abs	-	mug	obj
Estar en juego	joko	egon	-	-	-	-
Contraer matrimonio	[ez]	ezkondu	-	-	-	-

2. taula: Euskaratze-prozesurako informazioa jasotzeko taulen adibide bat.

Oraindik ez dugu euskarazko ordainei dagokien informazioa Matxinen integratu, eta, beraz, ez dakigu zehazki zenbateko ekarpena egingo duten datu horiek guztiek sisteman, baina gure susmoa da itzulpenen kalitatea nabarmen hobetuko dela aztertutako UFak dauzkaten esaldietan.

Konbitzul datu-basea

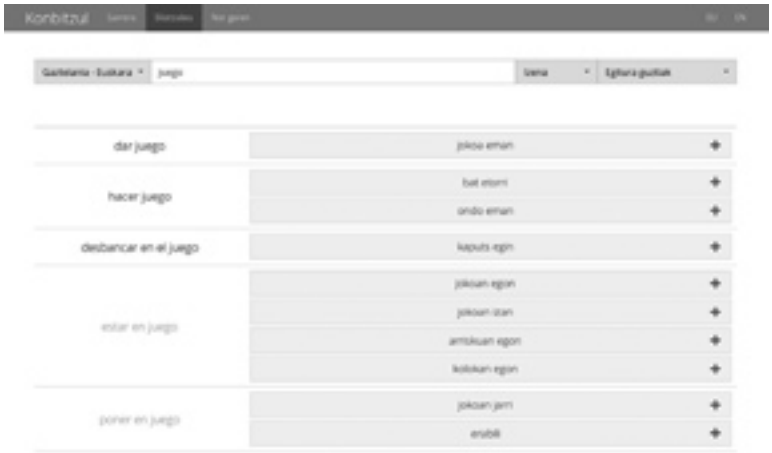
Sortu berri dugun datu-baseak Konbitzul du izena, eta artikulu honetan azaldutako ikerketalanean aztertzen ari garen informazioa jasotzen du. Datu-base publikoa da, http://ixa2.si.ehu.eus/konbitzul helbidean dago, eta edonork erraz erabiltzeko moduko interfazea du.

2. irudia: Konbitzul bilatzailearen itxura.

Bilaketak hiru irizpideren arabera egin daitezke:

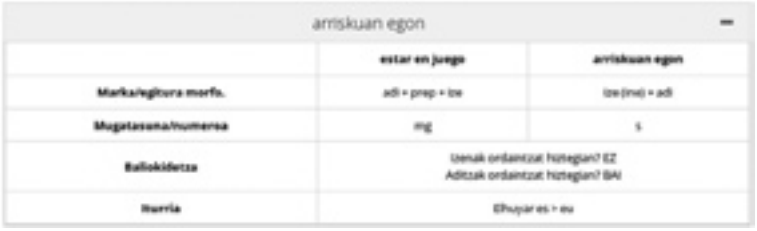
- Hizkuntza-norantza: gaztelaniatik euskarara ala euskaratik gaztelaniara
- Zer bilatu nahi den: aditza, izena, edo konbinazio osoa
- Egitura morfologikoa: euskarazko izenen kasu- edo postposizio-marka, edo gaztelaniazko konbinazioen egitura

Idatzitakoa datu-basean bilatzen da, eta bilaketarekin bat datozen emaitzen zerrenda erakusten da. Esate baterako, gaztelaniazko *juego* izena bilatuta, honako hau agertzen da:



3. irudia: *jugar* izena bilatuta Konbitzulek erakusten duena.

Bigarren irudian ikus daitekeenez, datu-basean *juego* izena duten bost konbinazio aurkitu dira: *dar juego, hacer juego, desbancar en el juego, estar en juego* eta *poner en juego.* Euskarazko ordainak sarreren eskuinaldean agertzen dira, eta horietako bakoitzak [+] ikur bat du aldamenean, informazio linguistikoa erakusteko edo ezkutatzeko. *Arriskuan egon* konbinazioaren gainean klikatuz gero, adibidez, hirugarren irudiko taula agertzen da.



4. irudia: *estar en juego – arriskuan egon* bikotearen informazio linguistikoa.

Horrez gain, gaztelaniazko sarrera batzuek ere, urdinez agertzen direnek, aukera ematen dute gainean klikatu eta informazio sakonagoa ikusteko. Hona hemen *estar en juego* konbinazioaren taula:

Sailkapen lexiko-semantikoa	figuratiboa
Sailkapen morfosintaktikoa	erdi-finkoa
Funtzio sintaktikoa	mod
IIn modifikatzailerik?	ez
IIS-adi bereizterik?	bai
Hitz-hurrenkera	finkoa
Egitura	estar en - juego
Preposizioa	en
Determinatzailea	-
Numeroa	s
Mugatasuna	mg

5. irudia: *Estar en juego* konbinazioari buruzko informazio sintaktikoa.

Ordainetan klikatuta erakusten den informazioa gaztelania-euskarazko bikoteari dagokio, eta artikulu honetako 2. atalean azaldu dugun azterketatik aterea da. Gaztelaniazko sarreretan klikatuta agertzen dena, berriz, informazio sintaktikoa da batez ere, konbinazioen detekziorako erabiliko duguna (hirugarren ataleko 1. azpiatala). Euskaratze-prozesuari buruzko datuak ere (hirugarren ataleko 2. azpiatala) laster argitaratuko ditugu sarean, Matxinen integratu eta proba batzuk egin ondoren.

Ondorioak eta etorkizuneko lanak

Gure ustez, hizkuntza-aplikazio informatikoek behar-beharrezkoa dute informazio linguistikoa, inoiz UFak ondo prozesatuko badituzte. Horregatik ari gara izen + aditz konbinazioen zenbait ezaugarri linguistiko aztertzen, Matxin itzultzaile automatikoa gai izan dadin UF horiek hobeto itzultzeko.

Lehenik eta behin, era horretako konbinazioek gaztelaniaren eta euskararen artean dituzten desberdintasunak gutxi-gorabehera ezagutzeko, Elhuyar hiztegiko izen + aditz konbinazioak aztertu ditugu. Espero genuen bezala, ikusi dugu oso konbinazio gutxi (% 10 inguru) itzultzen direla hitzez hitz hizkuntza batetik bestera (bigarren atala).

Ondoren, Matxin itzultzaile automatikoari begira, batetik, gaztelaniazko UFak automatikoki ezagutzeko lagungarriak diren zenbait ezaugarriri begiratu diegu, eta, bestetik, konbinazio horiek euskaratzeko baliagarriak izan litezkeen beste zenbaiti (hirugarren atala). Egindako esperimentuen arabera, gaztelaniazko UFen detekzioa asko hobetzen da gehitutako datuei esker, lehen ezagutzen ziren konbinazioak baino % 30 bat gehiago ezagutzen baitira orain. Gainera, informazio linguistiko hori guztia Konbitzul datu-base publikoan jarri dugu, edozein erabiltzaileren eskura (laugarren atala).

Aurrera begira, hainbat gauza egiteko asmotan gara: hasteko, euskaratze-prozesurako aztertu dugun informazioa Matxinen integratu, eta datu guztiak Konbitzulera gehitu nahi ditugu; bigarrenik, hasiak gara orain arte egindakoa ingelesa-euskara hizkuntza-bikoterako ere aztertzen, eta, hemendik pixka batera, informazio hori ere eskuragarri jarri nahi genuke datu-basean; eta, hirugarrenik, interesgarria litzateke informazio semantikoa prozesu guztian nola erabili ere aztertzea, batez ere adiera bat baino gehiago izan ditzaketen hitz-konbinazio batzuk automatikoki nola desanbigua litezkeen ikusteko.

Eskerrak

Lan hau Ekonomia eta Lehiakortasun Ministerioak Uxoa Iñurrietari emandako doktoretza aurreko diru-laguntzari esker egin dugu (BES-2013-066372), Euskal Herriko Unibertsitateko IXA ikerketa-taldean, SKATeR proiektuaren barruan (TIN2012-38584-C06-02). Eskerrak eman nahi dizkiegu Elhuyar Fundazioari eta, bereziki, Antton Gurrutxagari, haiei esker lortu baitugu lan honetarako behar genuen materiala.

BIBLIOGRAFIA

Corpas, G. (1997). *Manual de fraseología española*. Madril: Editorial Gredos.

Garate, G. (1998). 27173 *Atsotitzak, Refranes, Proverbes, Probervia.* Bilbao Bizkaia Kutxa Fundazioa.

Gurrutxaga, A. (2015). *Idiomatikotasunaren karakterizazio automatikoa: izena+aditza konbinazioak.* Euskal Herriko Unibertsitatea.

Howarth, P. A. (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making*, 75. lib., Walter de Gruyter.

Iñurrieta, U.; Aduriz I.; Díaz De Ilarraza, Arantza; Labaka, G. eta Sarasola, K. (2014). «Izen + aditz konbinazioen azterketa elebiduna, hizkuntza-aplikazio aurreratuei begira». *Linguamática*, *6* (2), 45.-55. or.

Lakarra, J. (1996). *Refranes y sentencias (1596) ikerketak eta edizioa.* Bilbo: Euskararen Lekukoak 19, Euskaltzaindia.

Mayor, A.; Alegria, I.; Díaz De Ilarraza, A.; Labaka, G.; Lersundi, M. eta Sarasola, K. (2009). «Matxin, euskararako lehenengo itzultzaile automatikoa», *Senez*, 37, 197.-220. or.

Mocoroa, J. M. (1990). *Ortik eta emendik. Repertorio de locuciones del habla popular vasca*, Bilbo: Labayru, Kultura eta Turismo Saila, Eusko Jaurlaritza.

Sag, I. A.; Baldwin, T.; Bond, F.; Copestake, A. eta Flickinger, D. (2002). «Multiword expressions: A pain in the neck for NLP», in *Computational Linguistics and Intelligent Text Processing*, Berlin: Springer, 1.-15. or.

Sanz, Z. (2014). *Unitate Fraseologikoen itzulpena: alemana-euskara. Literatur testuen corpusean oinarritutako analisia*, Euskal Herriko Unibertsitatea.

Sarasola, I. (1997). *Euskara batuaren ajeak*. Irun: Alberdania.

Urizar, R. (2012). *Euskal lokuzioen tratamendu konputazionala*, Euskal Herriko Unibertsitatea.

Zabala, I. (2004). «Los predicados complejos en vasco», in *Las fronteras de la composición en lenguas románicas y en vasco*, Deustuko Unibertsitatea, 445.-534. or.

Resumen

Las Unidades Fraseológicas son combinaciones que varían mucho de un idioma a otro y resultan problemáticas para disciplinas como la traducción, el aprendizaje de idiomas extranjeros o el procesamiento del lenguaje natural. En este artículo, se empieza por aclarar algunos conceptos relacionados con la fraseología, para pasar a dar algunos datos estadísticos sobre los cambios morfológicos que ocurren al traducir combinaciones del tipo verbo+sustantivo entre el euskera y el castellano. Asimismo, se explica por qué las Unidades Fraseológicas suponen un problema para el traductor automático Matxin, así como el estudio lingüístico que estamos llevando a cabo para intentar solucionar dichos problemas. Para terminar, se presenta la base de datos Konbitzul, que pone a disposición de los usuarios toda la información obtenida de este análisis.

Résumé

Les Unités Phraséologiques varient considérablement d'une langue à l'autre et elles constituent une grande difficulté pour la traduction, l'apprentissage de langues étrangères et le traitement du langage naturel. Dans cet article, tout d'abord, nous essaierons d'expliquer certains concepts liés à la phraséologie. Nous offrirons ensuite des données statistiques sur les variations morphologiques qui se produisent lorsqu'on traduit des combinaisons du type verbe + nom de l'espagnol en euskara et vice-versa. Nous expliquerons ensuite les raisons pour lesquelles les Unités Phraséologiques sont un problème pour le traducteur automatique Matxin et nous parlerons également de l'étude linguistique que nous menons actuellement afin de résoudre ces problèmes. Pour finir, nous présenterons la base de données Konbitzul, qui met à la disposition des usagers toute l'information recueillie dans cette étude.

Abstract

Phraseological Units are combinations that vary greatly from one language to another, and are therefore problematic for disciplines such as translation and second language learning, as well as for natural language processing. This article starts by clarifying several phraseology-related concepts, then introduces statistical data on the morphological changes that occur when verb+noun combinations are translated from Spanish into Basque and vice versa. Additionally, it explains why phraseological units pose problems for the automatic translator Matxin, and presents the linguistic study we are conducting in order to solve those problems. Finally, it introduces the Konbitzul database, which makes all the information gathered from this analysis available to users.