I. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

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

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

GIZA ZIENTZIAK

***Konbitzul:* euskarazko eta gaztelaniazko izen+aditz konbinazioen datu-basea,**

Uxoa Iñurrieta

32-38 or.
https://dx.doi.org/10.26876/ikergazte.i.03

ANTOLATZAILEA: udako euskal unibertsitatea
BABESLEAK: Eusko Jaurlaritza Gobierno Vasco; BFA DFB Bizkaiko Foru Aldundia Diputación Foral de Bizkaia; eman ta zabal zazu UPV EHU
LAGUNTZAILEAK: Deusto Universidad de Deusto Deustuko Unibertsitatea; Euskal Herria Udalbiltza; Mondragon Unibertsitatea; upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa

Konbitzul: euskarazko eta gaztelaniazko izen+aditz konbinazioen datu-basea

Uxoa Iñurrieta*

IXA taldea, Euskal Herriko Unibertsitatea
usoa.inurrieta@ehu.eus

Laburpena

Konbitzul datu-basea azterketa linguistiko baten emaitza da, eta euskarazko eta gaztelaniazko izen+aditz konbinazioak biltzen ditu, euren ordainekin eta beste hainbat datu linguistikorekin batera. Lan honek, batetik, azterketa linguistikoaren berri ematen du, eta, bestetik, datu-basean bilaketak egiteko sortu dugun interfaze publikoaren ezaugarriak azaltzen ditu.

Hitz gakoak: Hizkuntzaren Prozesamendua, euskarazko hitz-konbinazioak, gaztelaniazko hitz-konbinazioak, fraseologia, itzulpengintza

Abstract

Konbitzul is a database of Basque and Spanish noun+verb combinations. It is now available for public use, and it allows users to search for the appropriate translation of a given combination, along with other linguistic data. This paper describes the in-depth linguistic analysis we undertook to achieve this, and presents the database and the public interface.

Keywords: Natural Language Processing, word combinations in Basque, word combinations in Spanish, phraseology, translation

1 Sarrera eta motibazioa

Euskara ondo ezagutzen duen edonork gehiegi pentsatu gabe erabiltzen ditu *hitz egin* edo *falta izan* bezalako egiturak, aditz soilak balira bezala, berez izen batez eta aditz batez osatuta daudela ohartu gabe. *Min hartu* eta *min eman*en gisako lokuzioak ere ez zaizkio batere arrotzak egiten, eta normalean ez du *min sentitu* edo *min sortu* bezalakorik esaten, horiek ere ideia berbera adierazteko balio lezaketen arren. Era berean, norbait *adarra jotzen* ari zaiola entzutean, euskaldunak berehala ulertzen du txantxetan ari zaizkiola, eta ez zaio burura etortzen ez *adar*rik eta ez *jotze*rik.

Konbinazio horiek guztiak, ordea, nahiko bereziak dira, ez baitituzte hizkuntzaren ohiko arauak betetzen, osaeragatik, esaldian duten portaeragatik, edo esanahia ez delako gardena. Gainera, horrelako egiturak beste hizkuntzetara eramatea ere ez da lan erraza, hitzez hitzeko itzulpenak ez baitira egokiak izaten. Har ditzagun, adibidez, aurreko paragrafoan aipatutakoak:

(1) *Hitz egin* eta *falta izan*en ordainak aditz soilak dira gaztelaniaz: *hablar* eta *faltar/carecer* (eta ez **hacer palabra* eta **ser/tener falta/carencia*).

(2) Euskaraz *min hartu* eta *min eman* bi aditz desberdinekin osatzen badira ere, gaztelaniaz aditz berbera erabiltzen da bietarako, eta ez da euskarazkoen baliokidea: *hacer(se) daño* (eta ez **coger/dar daño*).

(3) *Adarra jotzea*, gehien-gehienetan, ez da *tocar el cuerno* gaztelaniaz, baizik eta *tomar el pelo*.

Hitz-konbinazioak, beraz, egitura konplexuak dira, eta zailtasun handiak ematen dituzte, batez ere ikuspuntu elebidunetik. IXA ikerketa-taldean, euskarazko eta gaztelaniazko izen+aditz konbinazioen

* Uxoa Iñurrietaren doktoretza-tesiaren zuzendariak Itziar Aduriz eta Kepa Sarasola dira, eta Arantza Díaz de Ilarrazak eta Gorka Labakak ere lan honetan parte hartu dute.

itzulpena aztertzen ari gara, eta datu-base bat sortu dugu gure azterketa linguistikotik lortutako emaitzekin. *Konbitzul*[1] bilatzaile publikoaren bidez, edozein erabiltzailek du aukera datu-base horretan bildutako konbinazioak, konbinazioen ordainak, eta horiei guztiei buruzko informazio linguistikoa eskuratzeko.

Gure ikerketa-lanaren helburu nagusia hitz-konbinazioak aztertu eta haien prozesamendu konputazionalerako oinarri linguistikoak jartzea da, hizkuntza-aplikazio aurreratuetan arazo handiak sortzen baitituzte. Horixe gertatzen da, adibidez, itzulpen automatikoan; izan ere, horrelako egiturak ondo itzultzea zaila bada hiztunontzat, are gehiago ordenagailu batentzat, hitzei ohiko ordainak ematen baitizkie defektuz eta, hortaz, tratamendu berezirik egin ezean, itzulpen literal traketsak sortzen baititu. Hona, esaterako, Matxin[2] itzultzaile automatikoaren bi emaitza:

(4) Gaztelaniazko itzulgaia: *La pareja contrajo matrimonio.*
Euskarazko itzulpen automatikoa: *Bikotea *ezkontza uzkurtu zen.*
Itzulpen zuzena: *Bikotea ezkondu egin zen.*

(5) Gaztelaniazko itzulgaia: *El plazo de inscripción ha vencido.*
Euskarazko itzulpen automatikoa: *Inskripzioko *epea garaitu du.*
Itzulpen zuzena: *Izen emateko epea amaitu da.*

Lan honetan, gaia bere testuinguruan kokatu ostean (2. atala), gure azterketa linguistikoa zertan datzan eta orain arte zer ondorio nagusi atera ditugun azalduko dugu (3.1. atala), bai eta *Konbitzul* datu-baseak zer ezaugarri dituen ere (3.2. atala).

2 Hitz-konbinazioen prozesamendu konputazionala: arloaren egoera

Hitz-konbinazioak hizkuntza baten fraseologiaren parte direla esan ohi da. Ematen diren definizioetan erabateko adostasunik ez dagoen arren, egile gehienak bat datoz konbinazioen ezaugarri nagusiak aipatzean (Sanz Villar, 2011; Baldwin eta Kim): hitz batez baino gehiagoz osatutako unitateak dira, esaldian portaera sintaktiko berezia izaten dute sarri, eta hitz-segida osoaren esanahiak ez du zertan osagaien esanahien batura izan. Izan ere, askotan, osagai guztiak ezagututa ere, ez da erraza konbinazio batek zer esan nahi duen ulertzea. Horrez gain, testuinguruak ere eragina izan dezake konbinazio batzuen erabileran, testu-alor batetik bestera duten agerpen-maiztasuna aldakorra izaten baita. Hortaz, hitz-segida bat hitz-konbinaziotzat hartzen da, baldin eta konbinazioko osagaien informazioa ez bada nahikoa konbinazio osoaren ezaugarri sintaktiko, distribuzional edo semantikoak jakiteko (Silberztein, 1990).

Ikerketa linguistiko gehienetan, konbinazioen berezitasunak aztertu eta ezaugarri morfosintaktiko eta semantikoen araberako multzoak egiten dira. Rafel (2004) eta Zabala (2004) lanetan, adibidez, aditz-egitura konplexuetako osagaiak aztertzen dira, eta garrantzi berezia ematen zaie esanahi "osorik" gabeko aditzei, hots, aditz arinei. Rodríguez eta García Murgak (2003) ere *egin* astuna (*ohea egin*) eta arina (*solas egin*) bereizten dituzte, bigarrenarekin osatutako egiturak idiomatikoak direla proposatzeko. Gainera, lan horretan bertan aipatzen denez, predikatu mota horrek "sintaxilarien arreta deitu izan du, eta bereziki ergatibitatearekin erlazionaturiko eztabaida asko sorrarazi ditu"; horren adierazgarri dira, esaterako, Levin (1983) eta Berro (2010).

Sailkapenei dagokienez, berriz, irizpide bat baino gehiago erabili izan da konbinazioak multzokatzeko. Batzuek idiomatikotasuna hartu izan dute ardatz (Howarth, 1998); beste batzuek finkapen sintaktikoari eman diote garrantzia (Corpas Pastor, 2001); eta beste batzuek, berriz, konposizionaltasunaren arabera mailakatu dituzte hitz-konbinazioak (Sag *et al.*, 2002).

Bestalde, esparru aplikatuagoetarako eginiko lanen artean, indar berezia hartu dute azkenaldian Hizkuntzaren Prozesamendura bideratutakoek (Seretan, 2013; Alonso Ramos, 1995). Asko eta asko hitz-konbinazioak testu-corpusetan automatikoki identifikatzeko egiten dira, gerora konbinazio horiek erauzi eta, adibidez, hiztegigintzan erabili ahal izateko. Beste esperimentu batzuetan, ostera, hitz-konbinazioen inguruko informazioa askotariko hizkuntza-aplikazioetan txertatzen saiatu izan dira, itzultzaile automatikoetan kasu (Wehrli *et al.*, 2009).

[1] `www.ixa2.si.ehu.es/konbitzul`
[2] `www.opentrad.com`

Azken urteotan, euskaraz ere argitaratu da hitz-konbinazioen tratamendu konputazionalaz diharduen lanik, eta, horien artean, aipagarriak dira bi doktoretza-tesi: Urizarrena (2012) batetik, eta Gurrutxagarena (2014) bestetik. Lehenak euskarazko lokuzioen azterketa linguistiko xehea egin eta haien tratamendu konputazionalerako oinarriak jartzen ditu. Bigarrenean, berriz, euskarazko izen+aditz konbinazioak automatikoki erauzteko eta sailkatzeko sistema bat aurkezten da.

Ildo elebidunetik, ordea, Sanz Villarrenaz (2011) gain ez da ia lanik egin gurean hitz-konbinazioen inguruan, eta are gutxiago prozesamendu konputazionalari dagokionez. Beraz, hutsune hori betetzera dator ikerketa-lan hau, beti ere ildo elebakarretik egindakoa abiapuntutzat hartuta.

3 Azterketa linguistikoa eta datu-basearen sorrera

Sarreran esan bezala, *Konbitzul* datu-basea azterketa linguistiko baten emaitza da. Datozen azpiataletan azalduko dugu zer aztertu dugun zehazki (3.1), bai eta datu-baseak eta interfazeak zer ezaugarri dituzten ere (3.2).

3.1 Izen+aditz konbinazio-itzulpenen azterketa linguistikoa

Gure ikerketa-lana corpus paraleloetan oinarritzeko asmoa izan dugu hasieratik, testu errealetan erabiltzen diren konbinazioak erauzteko iturririk egokiena direlakoan. Hala ere, beste ezertan hasi aurretik, interesgarria iruditu zaigu hiztegi elebidunetan zer-nolako konbinazioak jasotzen diren aztertzea ere, eta horixe izan da gure lehen pausoa. Elhuyarren euskara-gaztelania eta gaztelania-euskara hiztegiak izan ditugu aztergai gure lehen lanean (3.1.1), eta, behin hori amaituta, corpusen gaineko azterketarekin hasi gara (3.1.2).

3.1.1 Hiztegien gaineko azterketa

Elhuyarren hiztegi elebidunetan, hitz bakarreko sarrera arruntez gain, hitz anitzeko hainbat konbinazio eta esapide ere jasotzen dira. Guk, multzo horretatik, izenez eta aditzez osatutakoak hartu ditugu gure azterketaren oinarritzat: euskarazko 2.954 konbinazio (dagozkien 6.392 ordainekin batera), eta gaztelaniazko 2.650 (dagozkien 6.587 baliokideekin).

Euskarazko konbinazio guztiak izen batez eta aditz batez osatuak dira, nahiz eta izenek askotariko kasu- eta postposizio-markak izan: *lan egin, suak hartu, berriketan aritu, burutik egon, hitzekoa izan...* Gaztelaniazkoetan, berriz, aditzaz eta izenaz gain, tartean preposizio eta determinatzaileak ere onartu ditugu, *meter baza* edo *tener afecto* bezalakoez gain honelako egiturak ere landu ahal izateko: *ser una pena, saber de memoria, dejar a un lado.* Azterketa eskuz egin dugu osorik, eta hainbat ataletan banatu dugu (zehaztasun gehiago eta emaitza guztiak Iñurrieta *et al.* (2014) lanean daude jasota).

Lehenik eta behin, hizkuntza bateko eta besteko konbinazioei begiratu diegu: osaera morfologikoari, eta aditz eta izen motei. Atera ditugun ondorioen artean, aipagarria da euskarazko izenen hiru laurdenek baino gehiagok absolutibo-marka daramatela (*denbora galdu, dei egin*) eta bestelako markak askoz ere gutxiago erabiltzen direla (*jokoan jarri, deabruak hartu, amuari lotu, leporaino egon*). Gaztelaniazko konbinazioen artean, berriz, ez dago halako alderik, nahiz eta gehien errepikatzen den egitura aditza+determinatzailea+izena izan (*dar un toque*). Bestalde, nabarmentzekoa da bai hizkuntza batean bai bestean gehien errepikatzen diren aditzak oso arruntak direla eta, gainera, asko baliokideak direla euren artean: *egin - hacer, izan - ser/estar/tener, eman - dar, hartu - tomar...*

Bigarrenik, ordainak aztertzeari ekin diogu. Euskarazko konbinazioen gaztelaniazko baliokideei begiratuta, atentzioa ematen du aditz soilen kopuruak, askoz ere altuagoa baita izenez eta aditzez osatutako konbinazioena baino, ia bikoitza. Hain zuzen ere, % 58,07 dira *cosechar* edo *trabajar* bezalakoak (*uzta bildu* eta *lan egin*en ordainak), eta % 30,85 bakarrik *hacer el paripé* edo *dejar a un lado*ren gisakoak (*itxura egin* eta *bazterrean utzi*renak). Gaztelaniatik euskararako zentzuan, ostera, ordainen ia erdiak (% 48,54) dira izen+aditz motakoak (*abrir los ojos - begiak ireki, cobrar auge - gorantz hasi*), eta aditz soilak (*dar a luz - erditu*) ez dira laurdenera ere heltzen (%23,53).

Hirugarrenik, beste konbinazio baten bidez itzultzen diren konbinazioak pixka bat xeheago aztertu nahi izan ditugu, eta atera dugu ondorio interesgarririk hemen ere. Hasteko, ikusi dugu badagoela nolabaiteko lotura bat gaztelaniazko preposizioen eta euskarazko kasu- eta postposizio-marken artean; izan ere, gaztelaniazko preposiziodun konbinazioen ordainetan, absolutiboa baino ohikoagoak dira bestelako

markak, eta kontrako zentzuan ere gauza bera. Hau da: gaztelaniatik euskarara, askoz ere ohikoagoak dira *comer con apetito - gogoz jan* bezalako baliokidetzak *pasarse de la raya - gehiegikeriak egin* bezalakoak baino, eta euskaratik gaztelaniara ere sarriago errepikatzen dira *harira etorri - venir al caso* eta gisakoak, *txiripaz gertatu - sonar la flauta* eta antzekoen aldean.

Bestalde, mugatasuna eta numeroa ez dira oso sarri gordetzen hizkuntza batetik bestera, euskaratik gaztelaniara izenak mugagabean daudenean izan ezik. Multzo horretan, konbinazioen ordain gehienak ere mugagabeak dira (% 80,72): *aurpegiz ezagutu - conocer de vista, tiro egin - abrir fuego.*

Bukatzeko, hizkuntza bateko eta besteko konbinazioen izenak izenekin eta aditzak aditzekin baliokide ote diren ere aztertu nahi izan dugu, eta, espero genuen bezala, ikusi dugu oso gutxitan gertatzen dela bai izena eta bai aditza ordaintzat jasota egotea hiztegian: *bakean utzi - dejar en paz* (*bake - paz* eta *utzi - dejar*). Gehienetan, osagaietako bat edo biak aldatu egiten dira hizkuntza batetik bestera: *zarata egin - armar bulla* (*zarata - bulla* bai, baina ez *egin - armar*).

Laburbilduz, azterketa honetatik guztitik sarreran baieztatzen genuena ondoriozta daiteke, alegia, konbinazioen itzulpena ez dela lan erraza, hitzez hitzeko itzulpenek, gehienetan, ez baitute balio xede-hizkuntzan testu txukun eta natural bat sortzeko.

3.1.2 Corpus paraleloen gaineko azterketa

Hiztegietatik lor daiteke hitz-konbinazioek hizkuntza batetik bestera duten aldakortasunaren inguruko ikuspegi orokor bat, baina konbinazio horien erabilera erreala aztertu nahi bada, testu-corpusetara jo beharra dago. Beraz, gure bigarren azterketan, lehen-lehenik, hiztegitik erauzitako konbinazioak corpus paralelo batekin alderatu nahi izan ditugu. Erabili dugun corpusak gaztelaniaz eta euskaraz parekatutako 491.853 esaldi biltzen ditu, askotariko iturrietatik jasoak.

Lehenago landuta genituen gaztelaniazko 2.650 konbinazioetatik, 200 aurkitu ditugu corpusean. Elhuyar hiztegian, 200 konbinazio horiek 385 ordain baino ez dituzte, eta corpusean, berriz, 1.641 modutara itzulita agertu dira; hortaz, corpusetik ateratako informazioak gure datu-basea ordain berriz aberasteko balio izan digu. Horrez gain, hiztegiko konbinazioen beste aldaera batzuk ere hauteman ahal izateko, izenen eta aditzen lemak baino ez ditugu bilatu. Esaterako, gure datu-basean *alzar la voz* genuen, baina ez *alzar una voz, alzar su voz...* Izenaren eta aditzaren lema bakarrik bilatuta, aldaera gehiago ere lortu ahal izan ditugu, eta, hala, hiztegietan jasotako 200 konbinazio horiez gain, beste 698 forma berri ere bildu ditugu.

Konbinazio guztiak agerpen-kopuruaren arabera ordenatu, eta 132 hautatu ditugu –bost agerpenetik gorakoak– euren ezaugarri linguistikoak aztertzeko. Oraingo honetan, itzulpen automatikorako erabilgarria izan litekeen informazioa bildu nahi izan dugu, eta, hortaz, gaztelaniazko konbinazioen aldakortasuna izan da gure aztergai nagusia. Izan ere, gaztelaniatik euskararako itzultzaile automatiko batentzat, oso garrantzitsua da sorburu-hizkuntzako konbinazioak testuetan nola erabiltzen diren jakitea, detekzio-prozesua ahalik eta ondoen egiteko. Honako ezaugarri hauei begiratu diegu:

- Badarama preposiziorik? Zein?
- Izenak har dezake determinatzailerik? Derrigorrezkoa da?
- Izen-sintagma mugatua ala mugagabea da? Bietara joan daiteke?
- Izen-sintagma singularra ala plurala da? Bietara joan daiteke?
- Izen-sintagmak har dezake modifikatzailerik? (adjektiborik, sintagma preposizionalik...)
- Joan daiteke ezer izen-sintagmaren eta aditzaren artean? (adberbiorik, beste sintagmaren bat...)
- Osagaien ordena alda daiteke?

Datu horiek kontuan hartuta, konbinazioak konposizionaltasunaren arabera multzokatu ditugu, Sag *et al.* (2002) lanean proposatzen den sailkapenean oinarrituta: finkoak (% 3,37), erdi-finkoak (% 70,79) eta libreak (% 25,84). Izan ere, multzo batean edo bestean egon, tratamendua ere era batekoa edo bestekoa izatea komeni da, detekzio-prozesuak ahalik eta emaitzarik onenak eman ditzan.

Konbinazio finkoak (*dar paso (a algo)* eta antzekoak) beti elkarrekin agertzen dira, ezin da tartean elementurik sartu, eta ordena eta forma berean egoten dira beti. Hortaz, multzo horretako konbinazioak hautemateko, nahikoa da hitz-segida finkoak bilatzea (aditzaren flexioa gorabehera).

Konbinazio erdi-finkoek, ordea, arazo gehiago ematen dituzte. Izan ere, oso sarri, elementuren bat agertzen da izen-sintagmaren eta aditzaren artean (6. adibidea), eta baliteke osagaien ordena ere aldakorra izatea (7. adibidea), esaldia pasiboan dagoenean edo galdera bat denean, adibidez. Askotan, gainera, konbinazioaren beraren forma ere aldatu egiten da (8. adibidea). Hori dela eta, konputazionalki tratatzeko zailenak multzo honetako konbinazioak dira.

(6) *Se cubrieron las plazas.*
Se cubrieron en seguida las plazas.

(7) *Se ha fijado un plazo.*
El plazo ha sido fijado.

(8) *Hay que dar más pasos.*
Hay que dar un paso más.
Hay que dar el siguiente paso.

Azkenik, *comer con apetito* bezalakoak **konbinazio libretzat** hartu ditugu, gure ustez hitz-segida ohikoak baitira, berezitasun morfologiko, sintaktiko edo semantikorik gabeak. Hortaz, teorian, ez dute tratamendu berezirik behar, hitz bakoitzari dagokion ordaina emanda itzulpen onargarriak lortu beharko bailirateke.

3.2 *Konbitzul* datu-basea eta interfazea

Azterketa linguistikoetan lortu dugun informazioa publiko egiteko, datu-base bat sortu eta interneten eskuragarri jarri dugu, euskaraz eta ingelesez, bilatzaile-itxurako interfaze baten bidez: *Konbitzul*.

Bilatu nahi den testua idazteko hutsuneaz gain, hiru aukera-leiho daude bilaketa-irizpideak zehazteko: batetik, hizkuntza-norantza (euskaratik gaztelaniara ala gaztelaniatik euskarara); bestetik, idazten den testuaren forma (konbinazio osoa, aditza edo izena); eta hirugarrenik, bilatu nahi den konbinazioaren egitura. *Asmatu* aditza bilatuta, adibidez, 1. irudiko emaitzak agertzen dira: hiru konbinazio eta hiru ordain.

1 Figure: Interfazearen itxura

Ordain bakoitzak [+] zeinu txiki bat dauka aldamenean. Gainean klik eginda, gure azterketetan analizatutako ezaugarri guztiak biltzen dituen koadrotxo bat agertzen da.

Adibidez, 2. irudian, *arnasa hartu* konbinazioa eta sei ordain ageri dira. *Tomar aire*ren koadrotxoa irekita dagoenez, datuok ikus daitezke: euskarazko izenak absolutibo-marka du; gaztelaniazko ordaina aditz batez eta izen batez osatuta dago; euskarazko izenaren mugatasuna zalantzazkoa da, eta gaztelaniazkoa mugagabea; izenak ez dira baliokideak Elhuyar hiztegiaren arabera, baina aditzak bai; konbinazio-parea Elhuyar gaztelania-euskara hiztegitik erauzi da.

2 Figure: Informazio linguistikoa *Konbitzul*en

4 Ondorioak eta etorkizuneko lanak

Konbitzul datu-basea azterketa linguistiko baten emaitza da, eta euskarazko eta gaztelaniazko 5.604 izen+aditz konbinaziori eta horien 12.979 ordaini buruzko informazioa biltzen du. Gure azterketak bi oinarri izan ditu orain arte: batetik, Elhuyar hiztegi elebidunetako hitz-konbinazioak landu ditugu, eta, bestetik, konbinazio horiek corpus paralelo batean bilatu ditugu, testu errealetan zenbat eta nola erabiltzen diren ikusteko.

Hiztegi elebidunen azterketan lortu ditugun emaitzek argi uzten dute izen+aditz konbinazioen itzulpena zeinen konplexua den eta, ondorioz, zenbateraino den beharrezkoa hizkuntza-aplikazio aurreratuetan tratamendu egoki bat ematea, itzulpen automatikoan adibidez. Hori dela eta, hurrengo azterketa itzultzaile automatiko baten emaitzak hobetzera bideratu dugu, gaztelaniazko konbinazioak ondo detektatzeko beharrezko diren ezaugarriei garrantzi berezia emanda eta behar duten tratamenduaren arabera multzokatuta.

Lortu dugun informazio linguistikoa itzultzaile automatikoetan integratzea izango da gure ondorengo pausoa, itzulpenen kalitatea zenbateraino hobetzen den jakiteko. Horrez gain, informazio sintaktiko-semantikoa erabilgarria izan ote litekeen ere aztertu nahi dugu, orain arteko lanarekin alderatu eta sistema hobetzeko metodorik baliagarriena zein den ikusteko.

Lan honetatik eratorritako emaitza guztiak eskuragarri daude sarean, eta edozein erabiltzailek du aukera aztertu ditugun hitz-konbinazioen gainean bilaketak egiteko. Izan ere, hitz-konbinazioak hain fenomeno linguistiko konplexua izanik, sortu dugun baliabidea erabilgarria izango delakoan gaude, Hizkuntzaren Prozesamenduan dabiltzanentzat ez ezik, baita bestelako hainbat erabiltzailerentzat ere, hizkuntzalariak, itzultzaileak eta euskara-ikasleak tarteko.

Eskerrak

Doktoretza aurreko ikerketa-lan hau Ekonomia eta Lehiakortasun Ministerioko diru-laguntza bati esker egin ahal izan dugu (BES-2013-066372), SKATeR (TIN2012-38584-C06-02) eta QTLeap (FP7-ICT-2013.4.1-610516) proiektuen barruan. Elhuyarren laguntza ere ezinbestekoa izan da, haiek eman baitigute lanaren abiapuntutzat erabili dugun materiala. Era berean, eskerrak eman nahi dizkiegu Ruben Urizarri eta Mikel Artetxeri ere, gurekin jardun baitute proiektu honetan, lehenak aholkulari-lanetan, eta bigarrenak bilaketak egiteko interfazea prestatzen.

Erreferentziak

Alonso Ramos, Margarita. 1995. Hacia una definición del concepto de colocación: de JR Firth a IA Mel'cuk.

Baldwin, Timothy, eta Su Nam Kim. Multiword expressions. *Handbook of Natural Language Processing, second edition* .

Berro, Ane. 2010. Unergative Predicates in Basque Varieties: Consequences for the Ergative Case Asignment. Master tesia. UPV-EHU.

Corpas Pastor, Gloria. 2001. La traducción de unidades fraseológicas: técnicas y estrategias. *La lingüística aplicada a finales del siglo XX. Ensayos y propuestas, Alcalá, Universidad de Alcalá* 2.779–787.

Gurrutxaga, Antton. 2014. Idiomatikotasunaren karakterizazio automatikoa: izena+aditza konbinazioak. Doktoretza tesia. UPV-EHU.

Howarth, Peter. 1998. Phraseology and second language proficiency. *Applied linguistics* 19.24–44.

Iñurrieta, Uxoa, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka, eta Kepa Sarasola. 2014. Izen+aditz konbinazioen azterketa elebiduna, hizkuntza-aplikazio aurreratuei begira. *Linguamática* 6.45–55.

Levin, Beth. 1983. On the Nature of Ergativity. Doktoretza tesia. MIT.

Rafel, Joan. 2004. Los predicados complejos en español. In *Las fronteras de la composición en lenguas románicas y en vasco*, 445–534. Servicio de Publicaciones.

Rodríguez, Sonia, eta Fernando García Murga. 2003. IZEN+EGIN predikatuak euskaraz. *Euskal Gramatikari eta literaturari buruzko Jardunaldiak XXI. mendearen atarian (I-II)* 1.417–436.

Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake, eta Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, 1–15. Springer.

Sanz Villar, Zuriñe. 2011. Alemanetik euskarara itzulitako unitate fraseologikoen azterketarako jarraibideak. *Senez: itzulpen aldizkaria* 41.125–139.

Seretan, Violeta. 2013. On Collocations and Their Interaction with Parsing and Translation. In *Informatics*, volume 1, 11–31. Multidisciplinary Digital Publishing Institute.

Silberztein, Max. 1990. Le dictionnaire électronique des mots composés. *Langue française* 71–83.

Urizar, Ruben. 2012. Euskal lokuzioen tratamendu konputazionala. Doktoretza tesia. UPV-EHU.

Wehrli, Eric, Violeta Seretan, Luka Nerima, eta Lorenza Russo. 2009. Collocations in a rule-based MT system: A case study evaluation of their translation adequacy. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*, 128–135.

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