Aditza+izena konbinazioen itzulpen automatikoa, arau linguistikoen bidez

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

IXA taldea, EHU
*Bartzelonako Unibertsitatea
usoa.inurrieta@ehu.eus, itziar.aduriz@ub.edu, a.diazdeilarraza@ehu.eus
gorka.labaka@ehu.eus, kepa.sarasola@ehu.eus

Laburpena

Lan honek aditza+izena motako Unitate Fraseologikoak (UF) eta itzulpen automatikoa ditu aztergai. Gaztelaniazko UF sorta bat eta haien euskarazko ordainak aztertu ondoren, azterketa horretatik ateratako informazio linguistikoa Matxin itzultzaile automatikoaren sisteman sartu dugu, itzulpen-kalitatean zer eragin duen ikusteko. Bai metrika automatikoek eta bai eskuzko ebaluazioak argi erakutsi dute informazio linguistikoak sistemaren emaitzak hobetzen dituela.

Hitz gakoak: Fraseologia, Hizkuntzalaritza Konputazionala, Unitate Fraseologikoak, izen+aditz konbinazioak, Itzulpen Automatikoa, euskara, gaztelania

Abstract

This paper presents an analysis of verb+noun Multiword Expressions (MWEs) in Machine Translation (MT). After analysing a number of Spanish MWEs and their Basque translations, linguistic data was added into an MT system, namely Matxin, to see to what extent it affected translation quality. Both automatic evaluation metrics and a human evaluation showed that MWE-specific information improves the system's results.

Keywords: *Phraseology, Computational Linguistics, Multiword Expressions, Verb+Noun Combina-tions, Machine Translation, Basque, Spanish*

1 Sarrera

Hizkuntza bat baino gehiago hitz egiten ditugunok aspaldi ikasia dugu dena ezin dela hitzez hitz itzuli hizkuntza batetik bestera. Hizkuntzaren arabera, ideiak era batera edo bestera adierazten dira, eta batean naturala dena, sarritan, oso arrotza –edo are ulergaitza– gerta daiteke beste batean.

(1) Euskaraz: arreta *jarri*
Gaztelaniaz: *prestar* atención (lit.: arreta *mailegatu*)
Frantsesez: *faire* attention (lit.: arreta *egin*)
Ingelesez: *pay* attention (lit.: arreta *ordaindu*)

Horixe gertatzen da hizkuntzek bere-bereak dituzten hitz-konbinazio batzuekin, Unitate Fraseologikoekin (UF). Hiztunok inongo arazorik gabe erabiltzen ditugu ondo ezagutzen ditugun hizkuntzetan, baina osaera aldetik badituzte berezitasunak: batzuetan, esanahia ez da barruko hitzen esanahien batura izaten (*hanka sartu*); beste batzuetan, ez dituzte hitz-konbinazio arrunten ezaugarri morfosintaktiko berberak (*lo egin*, eta ez **loa egin*); eta, beste askotan, hitz batekin beste hitz jakin batzuk bakarrik erabiltzen ditugu, nahiz eta esanahi bera adieraz lezaketen beste hainbat ezagutu (*ondorioak atera*, eta ez *ondorioak egin/erauzi...*).

Testuetan etengabe erabiltzen dira UFak, bai ahoz eta bai idatziz; atzerriko hizkuntzak ikastean edo hizkuntzak ordenagailu bidez prozesatzean, ordea, arazo-iturri izaten dira. Begira zer-nolako esaldiak sortzen dituzten itzultzaile automatikoek batzuetan, UFren bat itzultzeko eskatzen badiegu:

(2) Itzulgaia: El *plazo vence* la semana que viene.
Itzulpen automatikoa: *Epea* datorren astean *garaituko du*.
Itzulpen-proposamena: *Epea* datorren astean *amaituko da*.

(3) Itzulgaia: Las ingenieras *llevaron* el proyecto *a cabo*.
Itzulpen automatikoa: Ingeniariek proiektua *eraman zuten kabora*.
Itzulpen-proposamena: Ingeniariek proiektua *burutu zuten*.

Lan honetan, gaztelaniazko aditza+izena motako UFez (1-3 adibideak) eta haien euskararako itzulpenaz arituko gara. UFen eta itzulpen automatikoaren alorreko ikerketa zertan den labur azaldu ostean (2. atala), *Matxin* itzultzaile automatikoak nola funtzionatzen duen deskribatuko dugu, bai eta gaur egun UFak nola tratatzen dituen ere (3. atala). Gero, ataza horretan sortzen zaizkion arazoak konpontzeko zer egiten ari garen azalduko dugu (4. atala): itzulpen zuzenak sortzeko zer informazio linguistiko duen beharrezkoa gure ustez, eta informazio hori erabilita zer emaitza lortu ditugun orain arte.

2 Unitate Fraseologikoak eta itzulpen automatikoa: orain arteko ikerketa

UFak fenomeno arazotsua dira Hizkuntzaren Prozesamendurako; hitz batez baino gehiagoz osatuta daude, baina ez dira beti konposizionalak esanahiari dagokionez eta, hortaz, hitz-konbinazio osoa hartu behar da kontuan hizkuntza-tresna aurreratuek ondo prozesa ditzaten (Sag *et al.*, 2002). Halako hitz-konbinazioak testuetan identifikatze hutsa nahiko lan nekeza izaten da, askotan ezaugarri morfosintaktiko malguak izaten dituztelako eta, ondorioz, ez delako nahikoa hiztegietan begiratu eta hitz batez baino gehiagoz osatutako sarrerak testuetan hitz-segida finkoak balira bezala bilatzea. UF batzuk –artikulu honetan hizpide izango ditugun aditza+izena konbinazioak, kasu– beste batzuk baino malguagoak izaten dira morfosintaktikoki (Iñurrieta *et al.*, 2016b), eta horrek bereziki zailtzen du haien prozesamendua.

Era askotara sailkatu izan badira ere, esaldien parte diren UFen artean[1], bi mota nagusi bereizi ohi dira: lokuzioak eta kolokazioak (Pastor, 1996; Urizar, 2012). Lokuzioetan, konbinazio osoaren esanahia ez da barruko hitzen esanahien batura izaten, opakoagoak edo gardenagoak izan daitezkeen arren; multzo horretakoak dira, adibidez, *adarra jo* eta *burua hautsi*. Kolokazioak, berriz, gardenak izaten dira, baina esanahiaren pisurik handiena bi hitzetako batek izan ohi du –Melc'uk-en arabera (1995), *oinarri*ak–, eta hitz horrek aukeratzen du zein beste hitzekin konbinatu daitekeen beste esanahi jakin bat adierazteko –zein *kolokatu*rekin, alegia–; hori gertatzen da, esate baterako, honako hauetan: *izerdia bota, pozez zoratu, interesa piztu*.

Bestalde, badira UF batzuk aditz *arin*ekin osatzen direnak[2]. Oso aditz arruntak izaten dira normalean –*egin, hartu, eman...*–, eta konbinazioen barruan esanahia "arindu"egiten zaie (Zabala, 2004), hau da, haien esanahiak pisua galtzen du, eta izenari aditz-izaera emateko balio dute nolabait. Hala, aditz horien ezaugarri semantikoak desberdinak izaten dira esanahi osoa dutenean eta arinak direnean (Rodríguez eta García Murga, 2003); *egin*en esanahia, adibidez, ez da berbera *ohea egin* (osoa) edo *solas egin* (arina) esatean. Halako konbinazioak oso ohikoak dira euskaraz, eta, besteak beste, inguruko hizkuntzetan aditz bakarrez adierazten diren ekintza asko adierazteko erabiltzen dira: *lan egin, lo hartu, hitz eman...* Zenbaitek hitz elkartuen multzoan sartu izan badituzte ere, joera nagusia halakoak hitz-konbinaziotzat hartzea da, elkarketatik bereiz (Azkarate, 1990).

Jackendoff-en arabera (1997), hiztun batek ezagutzen dituen UFen kopurua eta hitz soilena oso antzekoak dira, eta etengabe agertzen dira testuetan. Hiztegiek, ordea, oso UF gutxi jasotzen dituzte, eta lokuzioak izaten dira gehien-gehienak. Dena dela, dezente ikertu da azken hamarkadotan testuetatik UFak automatikoki erauzteko teknikak garatzeko (Gurrutxaga, 2015; Ramisch, 2015), eta UFak bereziki lantzen dituzten hiztegiak ere gehiagotu egin dira. Horren adierazgarri dira, adibidez, gaztelaniazko kolokazioen *DiCE* hiztegia (Alonso Ramos, 2006) eta ingelesezko kolokazioen *Oxford Collocations Dictionary* (Deuter, 2008).

Hizkuntza bakoitzak bere UFak izan ohi ditu, eta gehienetan ezin izaten dira hizkuntza batetik bestera hitzez hitz itzuli, batez ere sorburu- eta helburu-hizkuntzak oso tipologia desberdinetakoak direnean (Sanz, 2015; Iñurrieta *et al.*, 2016a). Hori hala izanik ere, sortzen diren baliabide gehienak elebakarrak

[1]Bestelakoak dira Enuntziatu Fraseologikoak (paremiak eta errutinazko formulak), esaldi osoak eratzen baitituzte (Urizar, 2012): *zozoak beleari, ipurbeltz; egun on; zer moduz?...*

[2]Aditz arindun konbinazioak, oro har, kolokazioen azpimultzotzat jotzen dira (Gurrutxaga, 2015), baina bada lokuziotzat hartzen dituenik ere (Urizar, 2012).

dira, hizkuntzak ikasten ari direnei edo idazketarako laguntza behar dutenei zuzenduak. UFen alderdi elebiduna, ordea, oso gutxi ikertu da orain arte, eta are gutxiago ikuspuntu konputazionaletik.

Gaur egun, bi motatako itzultzaile automatikoak dira ohikoenak: estatistikan oinarritutakoak eta erregeletan oinarritutakoak. Lehenengoek corpus paralelo erraldoietatik ikasten dutenez, zeharka bada ere, kontuan hartzen dute informazio fraseologikoa (Seretan, 2014). Euskara bezalako hizkuntza txikietarako, berriz, egokiagoak dira bigarren multzokoak, corpus paraleloak txiki samarrak baitira estatistikaren bidez kalitatezko itzulpen automatikoak lortzeko. Bestalde, erregeletan oinarritutako sistemek arau linguistikoak eta hiztegi elebidun orokorrak izaten dituzte oinarrian, eta horiek bakarrik ez dira nahikoa UFak zuzen itzultzeko; emaitza txukunak lortu ahal izateko, behar-beharrezkoa dute UFen inguruko informazio gehigarria (Wehrli *et al.*, 2009; Iñurrieta *et al.*, 2017).

3 Matxin itzultzaile automatikoa

Matxin erregeletan oinarritutako itzultzaile automatiko bat da (Mayor *et al.*, 2009), eta gaztelaniatik euskarara itzultzen du. Hiru fasetan egiten du lan:

1. **Analisia**. Gaztelaniazko itzulgaia automatikoki analizatzen da, *Freeling* tresna erabiliz (Padró eta Stanilovsky, 2012). Analisi horretatik, informazio morfologikoa eta sintaktikoa lortzen da: hitzak zer morfemaz osatuta dauden, nola multzokatzen diren[3], hitz-multzoek zer erlazio duten euren artean... Gaztelaniazko esaldiaren errepresentazio abstraktua lortzen da fase honetan.
2. **Transferentzia**. Analisian lortutako informazioa euskarara ekartzen da, baina oraindik ere era abstraktuan. Batetik, hiztegi elebidunen bidez, gaztelaniazko osagai lexikoei euskarazko ordainak ematen zaizkie, eta bestetik, gramatika-egitura ere transferitzen da. Bereziki gaztelaniatik euskarara itzultzean egin beharreko hainbat aldaketa ere fase honetan egiten dira, preposizio-postposizio egokitzapena kasu.
3. **Sorkuntza**. Euskarazko hitzak sintagmatan antolatu, eta sintagmei euskarazko hurrenkera ematen zaie lehenik. Gero, euskarazko elementu lexikoei eta haien informazio linguistikoari forma eman, eta euskarazko esaldia sortzen da. *Morfeus* prozesatzaile morfologikoa (Aduriz *et al.*, 1999) erabiltzen da azken ataza honetarako.

Hortaz, Matxinen sistemak bi oinarri nagusi ditu: arau linguistikoak eta hiztegiak. Bi oinarri horiek bakarrik, ordea, gehienetan ez dira nahikoa izaten Unitate Fraseologikoak zuzen itzultzeko. Hona hemen adibide batzuk:

(4) Itzulgaia: Deberán *prestar atención* al problema.
Itzulpen automatikoa: *Arreta mailegatu* behar izango diote arazoari.

(5) Itzulgaia: *Cubrieron* todas las *plazas*.
Itzulpen automatikoa: *Plaza* guztiak *estali* zituzten.

(6) Itzulgaia: Ella *perdió los estribos*.
Itzulpen automatikoa: Hark *nor bere onetik atera zuen*.

Batetik, sistemaren hiztegi elebiduna mugatua da, eta oso maiz erabiltzen diren UF asko eta asko ez ditu jasotzen, 4. adibideko *prestar atención* kasu. Bestetik, hiztegian dauden UFei ematen zaien tratamendua ere mugatua da, oso sinplea, eta ez dabil beti ondo.

Izan ere, gaur egun, hitz batez baino gehiagoz osatutako hiztegi-sarrerak hitz bakarra balira bezala prozesatzen dira, hitz-segida finkotzat. Hori dela-eta, *cubrir una plaza* kolokazioa hiztegian badago ere, 5. adibidean ez da ondo itzuli, ez delako hiztegi-sarreran bezala agertzen, baizik eta pluralean eta aurretik adjektibo bat eta artikulu zehaztu bat dituela. Horixe gertatzen da beste UF asko eta askorekin ere, morfosintaktikoki malguak direlako eta sistema ez dagoelako malgutasun hori tratatzeko prestatuta.

Bestelakoa da 6. adibidean dagoen arazoa. *Perder los estribos* UFa hiztegi-sarreran bezalaxe ageri da esaldian, eta sistemak ondo identifikatu du, baina ez du kontuan hartu euskarazko ordainaren informazio linguistikoa. Horregatik sortu du itzulpen okerra; *hura bere onetik atera zen* behar zuen euskaraz.

[3]Hizkuntzalaritza konputazionalean, sintaxi mailako hitz-multzoei *chunk* deitzen zaie; sintagmen antzekoak dira.

4 Aditza+izena konbinazioak itzultzeko metodo berria

Aurreko ataleko 4-6 adibideek erakutsi bezala, UFak ondo itzultzeko, Matxinek behar-beharrezkoa du informazio linguistikoa. Batetik, hiztegi-sarrera diren UFak itzulgaian identifikatu ahal izateko, nola erabiltzen diren jakin beharra du sistemak. Eta, bestetik, UF horien euskarazko ordaina zein den eta nola erabiltzen den ere zehaztu behar zaio.

Informazio hori guztia itzulpen-prozesuaren lehen eta bigarren faseetan gehitzea proposatzen dugu lan honetan: gaztelaniazko UFen identifikazioari dagokiona, analisian; eta euskarazko ordainei dagokiena, berriz, transferentzian. Datozen azpiataletan azalduko dugu bi ataza horietarako zer ezaugarri linguistikori begiratu diogun, bai eta zer emaitza lortu ditugun ere.

4.1 Gaztelaniazko aditza+izena UFen identifikazioa

Lehenago aipatu bezala, Matxinek, gaur egun, oso era sinplean tratatzen ditu UFak, eta, malgutasun morfosintaktikoa kontuan hartzen ez duenez, UFen aldaera asko ez ditu identifikatzen. Begira zer gertatzen den honako adibide honetan:

(7) Itzulgaia: *Estoy de acuerdo. Estoy* muy *de acuerdo.*
Itzulpen automatikoa: *Bat nator.* Oso *nago akordiotik.*

Hiztegian badagoenez *estar de acuerdo* sarrera, lehen esaldian ondo identifikatu da UFa, hitz guztiak jarraian, hurrenkera berean eta –aditzaren flexioa gorabehera– forma berean ageri direlako. Bigarrenean, aldiz, aditzaren eta preposizio-sintagmaren artean beste hitz bat ageri denez, ez da UFrik identifikatu eta, ondorioz, esaldia oso trakets ekarri da euskarara.

Hala ere, zabalegi jokatzea ere ez da komenigarria, posible baita UF ez diren hitz-konbinazio asko UFtzat hartzea. Adibidez, esaldian UFko izena eta aditza agertzen ote diren bakarrik begiratuta, sistemak ez luke bereiziko 8. adibideko bi esaldien arteko aldea.

(8) Iban *dando voces* por la calle. ('Kalean *oihuka ari* ziren')
Las *voces* le *dieron* una pista. ('*Ahotsek* arrasto bat *eman* zioten')

Beraz, beharrezkoa da malgutasun morfosintaktikoa kontuan hartzea, baina oreka bilatu behar da metodo malguegien eta zurrunegien artean. Horretarako, *Konbitzul* datu-basean[4], izenez eta aditzez osatutako UF sorta bati buruzko informazio linguistikoa gorde dugu. Honako ezaugarri hauek aztertu ditugu:

- Determinatzaileak izen-sintagman: ager daitezke? Beti dira zehaztuak/zehaztugabeak ala bietakoak izan daitezke?
- Izen-sintagmaren numeroa: beti da singularra/plurala ala batera zein bestera erabiltzen da?
- Modifikatzaileak izen-sintagman: ager daitezke? (adjektiboak etab.)
- Izen- edo preposizio-sintagmak betetzen duen funtzioa: objektua, subjektua ala modifikatzailea?
- Aditzaren eta izen- edo preposizio-sintagmaren bereizgarritasuna: ager daiteke beste hitzen bat bi elementuen artean? (adberbioak etab.)
- UFko osagaien hurrenkera: beti finkoa da ala alda daiteke? (galderetan etab.)

Ezaugarri horien arabera, UFak hiru multzotan sailkatzen dira datu-basean: finkoak[5], erdi-finkoak eta malguak. Multzo horien arabera, teknika bat edo beste aplikatzen da UFak identifikatu ahal izateko. Itzulgaia automatikoki analizatzean (3. atala), esaldietan UFrik badagoen begiratzen da, Konbitzulen laguntzaz:

- UFa erdi-finkoa denean, datu-basean zehaztutako ezaugarriei begiratzen zaie. Adibidez, *dar voces* konbinaziorako, Konbitzulek Matxini aginduko dio izen-sintagma pluralean dagoenean eta aditzaren objektua denean bakarrik hautemateko. Hala, *dar voces* UFa identifikatuko du honako esaldiotan:

[4] `http://ixa2.si.ehu.eus/konbitzul`

[5] Aditza+izena konbinazioak bereziki malguak izaten dira, eta ez dugu guztiz finkoa denik aurkitu landu ditugunen artean.

(9) Iban *dando voces* por la calle.
Daban siempre *voces* por la calle.
Con las *voces* que *daban*, se les oía desde la calle.

Baina ez besteotan:

(10) Fue él quien *dio* la *voz* a varios personajes televisivos.
Hace falta *dar* la *voz* de alarma.
El objetivo de la iniciativa es *dar voz* a los que no la tienen.

- Konbinazioa guztiz malgua denean, berriz, sistemak begiratzen du ea aditzak eta izenak erlazio zuzena duten analisi-zuhaitzean, hau da, ea izen-sintagma aditzaren subjektua, objektua edo modifikatzailea den. Hala, lehen ez bezala, *fijar un plazo* UFa ondo identifikatu ahal izango da esaldi hauetan guztietan:

(11) *Fijaron* el *plazo* de inscripción.
Mañana *fijarán* un nuevo *plazo*.
Los *plazos* deben ser *fijados* con antelación.

4.2 Aditza+izena UFen euskaratzea

Behin UFak identifikatutakoan, Matxini informazioa eman behar zaio euskarara nola ekarri behar diren jakin dezan. Batzuetan nahikoa da aditzari eta izenari zer ordain eman behar zaion zehaztea; beste batzuetan, informazio gramatikala da kontuan hartu beharrekoa; eta beste batzuetan, bai bata eta bai bestea.

Identifikaziorako informazioarekin egin bezala, itzulpenari dagokion informazioa ere Konbitzul datubasean gorde dugu. Honako ezaugarri hauei begiratu diegu:

- Izen-sintagmaren kasu- edo postposizio-markari
- Izen-sintagman ager daitezkeen determinatzaileei
- Izen-sintagmaren numeroari eta mugatasunari
- Izen-sintagmaren eta aditzaren arteko erlazioari
- UFtik kanpoko elementuen kasu- edo postposizio-markei

Informazio gramatikalik aldatu behar ez denean, nahikoa da UFko aditzari eta izenari zer ordain dagozkien zehaztea. Honako esaldi honetan, adibidez, aditzaren ordaina aldatu behar da: *esnatu*ren ordez, *piztu*.

(12) Itzulgaia: El tema *despertó interés*.
Itzulpen automatikoa: Gaiak *interesa esnatu* zuen.
Itzulpen-proposamena: Gaiak *interesa piztu* zuen.

Lexikoaren ordez gramatika denean tratatu beharrekoa, arau gehigarriak sartu behar dira sisteman. Esate baterako, 13. adibidean, ez da beharrezkoa lexikoa aldatzea, baina euskarazko izen-sintagmaren postposizio-marka –kasu honetan, instrumentala– zehaztu beharra dago, ez baita Matxinek gaztelaniazko preposizioari defektuz emango liokeena –kasu honetan, soziatiboa–.

(13) Itzulgaia: Lo *trata con respeto*.
Itzulpen automatikoa: *Errespetuarekin tratatzen* du.
Itzulpen-proposamena: *Errespetuz tratatzen* du.

Azkenik, badira UF batzuk bai tratamendu lexikala eta bai gramatikala behar dutenak, 14. adibidean gertatzen den bezala:

(14) Itzulgaia: La *echan en falta*.
Itzulpen automatikoa: *Faltan botatzen* dute.
Itzulpen-proposamena: Haren *falta sumatzen* dute.

Batetik, aditzari ez zaio ohiko ordainik eman behar *echar en falta* UFa itzultzeko: *bota*ren ordez, *sumatu*. Bestetik, preposizioari ez zaio postposiziorik eman behar ordaintzat, baizik eta kasu absolutiboari dagokion marka; eta, gainera, gaztelaniazko esaldian objektu zuzenaren funtzioa betetzen duena izen-sintagmaren modifikatzaile bihurtzen da euskaraz, eta genitibo-marka jarri behar zaio.

Informazio hori guztia arauen bidez gehitu dugu Matxinen sisteman, eta, jarraian, lagin batekin probak egin eta sistema ebaluatu dugu.

4.3 Emaitzak

Proposatzen dugun metodoa Matxinentzat lagungarria den ala ez jakiteko, linguistikoki aztertutako 92 UF hartu –gure datuen arabera gaztelaniazko testuetan maizen agertzen direnak–, eta haiei buruzko informazioa sartu dugu sisteman. Gaztelaniaren eta euskararen artean itzulitako testuak biltzen dituen corpus paralelo batetik, UFetako izena eta aditza barne hartzen dituzten esaldiak hautatu ditugu, eta 1.991 esaldi-pareko azpicorpus bat osatu dugu. Corpus hori oinarri hartuta, bi eratara ebaluatu dugu jatorrizko sistemaren eta berriaren arteko aldea: metrika automatikoak erabiliz eta eskuz.

4.3.1 Metrika automatikoak

Metrika bat baino gehiago dago itzultzaile automatikoen kalitatea neurtzeko. Guk honako hiru hauek aukeratu ditugu:

- **BLEU** (Papineni *et al.*, 2002), itzulpen automatikoaren alorreko ebaluazio-metrikarik ezagunena. Sistemaren erantzunak esaldika hartu, eta ereduzko itzulpenekin (corpus paralelokoekin) alderatzen ditu. Sorburu-hizkuntzako esaldiak zatikatzen joaten da –lehenengo, hitz bakarra; gero, bi hitzeko multzoak, etab.–, eta zati horiei corpusean eman zaizkien itzulpenak bilatzen ditu, aztertzeko zenbateraino diren antzekoak sistemaren itzulpenak eta eskuzkoak.
- **NIST** (Doddington, 2002). BLEUn oinarrituta dago, baina beste ezaugarri bat ere hartzen du kontuan: ebaluatzen dituen hitz multzoen maiztasuna. Hitz multzo jakin bat zenbat eta arraroagoa izan ereduzko corpusean, sistemaren itzulpena zuzena bada, orduan eta pisu handiagoa ematen dio, eta alderantziz –zenbat eta ohikoagoa izan, orduan eta pisu txikiagoa–.
- **TER** (Snover *et al.*, 2006). Metrika honek ere sistemaren emaitzak ereduzko itzulpenekin alderatzen ditu, baina beste era batera egiten ditu kalkuluak: itzulpen automatikotik eskuzkora iristeko egin beharreko moldaketa-kopuruaren arabera. Hortaz, BLEUn eta NISTen ez bezala, zenbat eta txikiagoa izan TERen balioa, itzulpen automatikoaren kalitatea orduan eta hobea dela esan nahi du.

Hiru ebaluazio-metriken emaitzak hobetzen dira UFei buruzko informazioa sisteman gehitu ondoren (ikus 1. taula), baina BLEU neurriak erakusten du alderik handiena bi sistemen artean, % 3,02koa (0,22 puntu).

Sistema	**BLEU**	**NIST**	**TER**
Matxin-UF	7,50	3,90	84,27
Matxin	7,28	3,88	84,36

1 Taula: BLEU, NIST eta TER emaitzak, Matxinen UFei buruzko ezagutza gehituta eta gehitu gabe

Dena dela, kontuan hartu behar da erabili dugun corpusa esperimentu honetarako sortu dugula espreski, landutako konbinazioen itzulpena aztertzeko. Corpus orokor handiago bat erabiliko bagenu, 92 UFek askoz ere agerpen gutxiago izango lituzkete, eta, ziur asko, metrika automatikoetan lortutako hobekuntza oso txikia –ia hautemanezina– izango litzateke.

4.3.2 Eskuzko ebaluazioa

Metrika automatikoak erabiltzeaz gain, eskuzko ebaluazio kualitatiboa ere egin dugu, eta hortik atera ditugu ondoriorik interesgarrienak. Bi sistemek desberdin itzulitako esaldi sorta erakusgarri bat hartu, eta hiru ebaluatzaileri eman diegu: (A) hizkuntzalari bati, (B) itzultzaile bati eta (C) euskaraz eta gaztelaniaz arazorik gabe egin arren hizkuntza-ikasketa bereziturik ez duen hiztun bati. Gaztelaniazko esaldi bakoitzarekin batera sistema baten eta bestearen itzulpenak erakutsi, eta aukeratzeko eskatu diegu: lehena hobea den, bigarrena hobea den, ala biak diren maila berekoak. Emaitzak 2. taulan jaso ditugu.

Ebaluatzaileen erantzunei begiratuta, hobekuntza nabaria da jatorrizko sistematik berrira. Baina beste zerbait ere iradokitzen dute: adituak direnentzat begi-bistakoak diren hobekuntzak ez direla hain

Sistema	A	B	C
Matxin-UF hobea	77,50%	77,50%	46,50%
Matxin hobea	8%	6,50%	40,50%
Maila berekoak	14,50%	16%	13%

2 Taula: Eskuzko ebaluazioaren emaitzak, anotatzailearen arabera

agerikoak hizkuntza-ikasketa bereziturik ez duten hiztunentzat. Izan ere, esaldi guztien % 43,52k kontraesanak sortu dituzte hiru ebaluatzaileen artean, baina horietako % 78,57tan C anotatzailea izan da esaldia desberdin ebaluatu duena, esaldi guztien % 33tan, alegia.

Dena dela, euskara normalizazio-prozesuan egonik, ez da harritzekoa UF batzuen aurrean hiztun guztiek sentipen berbera ez izatea. Esate baterako, kontraesan gehien sortu dituen konbinazioa *dar pasos* izan da: Matxinen jatorrizko sistemak *urratsak eman* itzultzen zuen, eta Konbitzuleko informazioa darabilenak, berriz, *pausoak eman*. Euskarazko tradizioan *urratsak egin* zein *pausoak eman* erabili izan dira, *urratsak eman*en agerpen bakanen bat ere badagoen arren –Orotariko Euskal Hiztegiaren arabera (Mitxelena, 1987), hegoaldeko autore moderno batzuen testuetan–. Gaur egungo testuetan begiratuta ere, lehen biak dira usuen agertzen direnak, baina hirugarrena ere nahiko sarri errepikatzen da. Beraz, ez da harritzekoa adituek *pausoak eman* aukeratu izana, baina ezta beste hiztun batzuei *urratsak eman* normal-normala iruditzea ere. Horrelakoetan ere hobekuntza badagoela uste dugu guk, nahiz eta hiztun guztientzat begi-bistakoa ez izan.

Beraz, oro har, sistema hobetu dela ulertu dugu honako kasuotan: hiru ebaluatzaileek sistema berriaren alde egin dutenean, bik berriaren alde egin eta hirugarrenak bi sistemak berdintzat jo dituenean, eta A eta B ebaluatzaileek sistema berriaren alde egin eta kontraesana C ebaluatzaileak sortu duenean. Horiek guztiak batuta, hobekuntza osoa % 78,6koa dela ondorioztatu daiteke eskuzko ebaluaziotik.

5 Ondorioak eta etorkizuneko lanak

Matxin itzultzaile automatikoak oso era sinplean tratatzen ditu UFak, eta horrek oso itzulpen traketsak sorrarazten dizkio sarri. Lan honetan, aditza+izena motako UFak automatikoki itzultzean sortzen diren zailtasunak aztertu ditugu, eta zailtasun horiei aurre egiteko metodo bat proposatu dugu.

Batetik, gaztelaniazko UFak identifikatzen laguntzeko, haien malgutasun morfosintaktikoa aztertu eta hainbat datu sartu ditugu sisteman: numeroa eta determinatzaileak zenbateraino diren aldakorrak, izensintagman modifikatzailerik sar ote daitekeen, etab. Bestetik, identifikatutako UFak euskarara ekartzeko, bi eratako informazioari begiratu diogu: lexikalari eta gramatikalari. Aztertutako datuak Matxinentzat lagungarriak ote ziren jakiteko, 92 UFri buruzko informazioa sartu dugu sisteman, eta 1.991 esaldiko corpusa itzuliarazi diogu. Metrika automatikoek zein eskuzko ebaluazioak argi erakutsi dute Matxinen itzulpenak hobetu egiten direla UFei buruzko informazio linguistikoa gehitzean: BLEUn lortutako marka % 3,02 igotzen da, eta giza ebaluatzaileen lanetik ondorioztatu daiteke hobekuntza % 78,6koa izan dela.

Aurrera begira, Konbitzul datu-basea elikatzen jarraitzea da gure asmo nagusia. Horretarako, UFak bereziki lantzen dituzten baliabideak eta corpus paraleloak ustiatuko ditugu, eta eskuz egin dugun azterketa linguistikoa era erdi-automatikoan egin ote daitekeen ere ikertuko dugu. Bestalde, interesgarria litzateke informazio semantikoa sakonago aztertzea ere, orain arte ezaugarri morfosintaktikoei eman baitiegu garrantzia batik bat.

Erreferentziak

Aduriz, Itziar, Eneko Agirre, Izaskun Aldezabal, Xabier Arregi, Jose Mari Arriola, Xabier Artola, Koldo Gojenola, Montse Maritxalar, Kepa Sarasola, eta Miriam Urkia. 1999. *MORFEUS: Euskararako analizatzaile morfosintaktikoa*. Barne-txostena, UPV/EHU/LSI/TR.

Alonso Ramos, Margarita. 2006. Glosas para las colocaciones en el diccionario de colocaciones del español. *Diccionario y Fraseología. ed. by M. Alonso Ramos* 59–88.

Azkarate, Miren. 1990. *Hitz elkartuak euskaraz*. Deustuko Unibertsitatea, Filosofi-Letren Fakultatea.

Deuter, Margaret. 2008. *Oxford Collocations Dictionary: for students of English*. Oxford University Press.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, 138–145. Morgan Kaufmann Publishers Inc.

Gurrutxaga, Antton. 2015. *Idiomatikotasunaren karakterizazio automatikoa: izena+aditza konbinazioak*. Doktorego-tesia, Informatika Fakultatea, UPV/EHU, Donostia.

Iñurrieta, Uxoa, , Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka, eta Kepa Sarasola. 2017. Rule-based translation of spanish verb+noun combinations into basque. In *Proceedings of the 13th Workshop on Multiword Expressions, EACL 2017*.

——, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka, eta Kepa Sarasola. 2016a. Izen+aditz konbinazioen itzulpenaz eta tratamendu konputazionalaz. *Senez, 47* 237–249.

——, Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola, Itziar Aduriz, eta John Carroll. 2016b. Using linguistic data for english and spanish verb-noun combination identification. In *The 26th International Conference on Computational Linguistics (COLING 2016): Technical Papers*, 857–867.

Jackendoff, Ray. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Mayor, Aingeru, Iñaki Alegria, Gorka Díaz de Ilarraza, Arantza adn Labaka, Mikel Lersundi, eta Kepa Sarasola. 2009. Matxin, euskararako lehenengo itzultzaile automatikoa. *Senez, 37* .

Melcuk, Igor. 1995. Phrasemes in language and phraseology in linguistics. *Idioms: Structural and psychological perspectives* 167–232.

Mitxelena, Koldo. 1987. Orotariko euskal hiztegia.

Padró, Lluís, eta Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, 2473–2479.

Papineni, Kishore, Salim Roukos, Todd Ward, eta Wei-Jing Zhug. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Pastor, Corpas. 1996. *Manual de fraseología española*. Editorial Gredos, Madrid.

Ramisch, Carlos. 2015. *Multiword Expressions Acquisition*. Springer.

Rodríguez, Sonya, eta Fernando García Murga. 2003. Izen+ egin predikatuak euskaraz. *Euskal gramatikari eta literaturari buruzko ikerketak XXI. mendearen atarian, Iker-14* 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 *International Conference on Intelligent Text Processing and Computational Linguistics*, 1–15. Springer.

Sanz, Zuriñe. 2015. *Unitate Fraseologikoen itzulpena: alemana-euskara*. Doktorego-tesia, Letren Fakultatea, UPV/EHU, Gasteiz.

Seretan, Violeta. 2014. On collocations and their interaction with parsing and translation. In *Informatics*, volume 1, 11–31. Multidisciplinary Digital Publishing Institute.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, eta John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

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

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.

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

Eskerrak eta oharrak

Lan hau Ekonomia eta Lehiakortasun Ministerioak Uxoa Iñurrietari emandako diru-laguntza bati esker egin dugu (BES-2013-066372), SKATeR (TIN2012-38584-C06-02), EXTRECM (TIN2013-46616-C2-1-R) eta TADEEP (TIN2015-70214-P) proiektuen barruan.