VI. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2025eko maiatzaren 28, 29 eta 30a
Bilbo, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

Aitortu-PartekatuBerdin 4.0

GIZA ZIENTZIAK ETA ARTEA

Sinonimoen erabilera albiste digitaletan: testu originalen eta itzulpen automatikoen arteko alderaketa

Amaia Solaun Martínez eta
Nora Aranberri Monasterio

133-139 or.
https://dx.doi.org/10.26876/ikergazte.vi.01.16

ANTOLATZAILEA
BABESLEAK
LAGUNTZAILEAK

Sinonimoen erabilera albiste digitaletan: testu originalen eta itzulpen automatikoen arteko alderaketa

Amaia Solaun Martínez, Nora Aranberri Monasterio
Hizkuntza Teknologiako Euskal Zentroa (HiTZ),
Euskal Herriko Unibertsitatea (UPV/EHU)
amaia.solaun@ehu.eus nora.aranberri@ehu.eus

Laburpena

Itzultzaile automatikoen hedatze azkarraren ondorioz inoiz baino harreman estuagoa dugu itzulpen automatikoekin. Egoera horrek euskararen erabileran zuzenean eragin dezake, teknologia honek hizkuntza aldaketak eragiteko potentziala baitu. Horregatik, lan honetan euskarazko albiste digital originalen eta itzulpen automatikoen arteko desberdintasunak ikertu ditugu. Zehazki, sinonimoen erabilera aztertu dugu metodo eta metrika automatikoak erabilita. Gure emaitzek aditzera ematen dute orobat sinonimoen erabilera oso antzekoa dela bi testu moten artean, baina, hala ere, giza testuek aniztasun handiagoa dutela, sinonimoen erabilera apur bat orekatuagoa baitute.

Hitz gakoak: itzulpen automatikoa, hizkuntza aniztasuna, sinonimia, hizkuntza gutxituak, euskara

Abstract

Due to the rapid expansion in the use of machine translation systems, we have a closer relation to its output than ever before. This could have direct implications in the use of the Basque language, as this technology has the potential to influence linguistic change. For this reason, in this work we address the differences between original and machine generated Basque online news. More precisely, we study the use of synonyms in each of the texts using automatic methods and metrics. Our findings reveal a notable degree of similarity overall between the two texts. However, we do notice a slightly more balanced and, therefore, more diverse use of synonyms in the human generated content.

Keywords: machine translation, language diversity, synonymy, minority languages, Basque

1. Sarrera eta motibazioa

Itzultzaile automatikoen kalitate eta irisgarritasunean emandako aurrerapausoen ondorioz, teknologia honen erabilera azkar zabaltzen ari da, hala esparru profesionalean nola erabiltzaile orokorren artean. Ondorioz, gero eta kontaktu handiagoa dugu itzulpen automatikoekin, baita hizlari gutxiagoko hizkuntzetan ere, besteak beste euskaran (Aranberri & Iñurrieta, 2024).

Testuinguru honetan, hainbat ikerketak azpimarratu dute itzulpen automatikoaren bidez sortatuko testuek ezaugarri bereizgarriak dituztela, eta giza itzulpenenetatik nahiz testu originaletatik desberdinak direla. Esate baterako, sorburu testuko interferentzia handiagoa (Castilho et al., 2019) eta hizkuntza aniztasun txikiagoa izan dezaketela behatu da (Vanmassenhove et al., 2019, 2021). Ezaugarri horiek, ziurrenik, algoritmoek hizkuntzaren konplexutasuna erreproduzitzeko dituzten zailtasunen ondorio dira, eta, beraz, gaitzesgarritzat har daitezke, batez ere hizkuntzaren erabilera oker edo ezohikoak ezaugarri horien parte direnean. Horrek guztiak itzulpen automatikoaren masifikazioaren inguruan zalantzak planteatzen ditu, ez baitago argi nola eragin dezakeen hizkuntzaren erabileran, eta bereziki nola eragin dezakeen hizkuntza gutxituetan, hala nola euskaran.

Inguruko hizkuntza nagusituekin alderatuta, euskara posizio zaurgarriago batean dago. Izan ere, testuinguru diglosiko batean egoteaz gain, normalizazio prozesu batean ere murgilduta dago, duela gutxi estandarizatu baita (Zabaleta, 2019). Horrez gain, teknologiaren ikuspuntutik, euskara baliabide urriko hizkuntza da (Hernáez et al., 2012; Sarasola et al., 2022), eta, horren ondorioz, itzultzaile automatikoen kalitatea baxuagoa da baliabide ugariko hizkuntzekin alderatuta. Dena den, arestian egindako inkesta baten arabera, itzultzaile automatikoen erabilera gero eta handiagoa da euskal komunitatean, eta ez da espero etorkizunean hazkunde hori gelditukо denik (Aranberri & Iñurrieta, 2024). Horrenbestez, eta euskararen posizio zaurgarria kontuan hartuta, posible da hizkuntza gero eta sentikorragoa bihurtzea itzulpen automatikoaren

eraginarekiko. Horregatik, ezinbestekoa da itzulpen automatikoen eta gizakiek sortutako edukiaren arteko desberdintasunak sakontasunez aztertzea.

Ikerlan honek urrats bat ematen du norabide horretan, eta itzulpen automatikoen azterketa bat proposatzen du, teknologia honek aniztasun lexikoan izan ditzakeen efektuak aztertzeko. Zehazki, azterlan honetan testu originalen eta itzulpen automatikoen sinonimoen erabilera alderatzen da. Horretarako, EITB corpus konparagarria erabiltzen dugu (Etchegoyhen et al., 2020), euskarazko eta gaztelaniazko berriak biltzen dituena, eta gaztelaniatik euskarara itzultzen dugu ELIA itzultzailea erabiliz. Azkenik, lortutako itzulpenak euskarazko albiste originalekin alderatzen ditugu.

2. Arloaren egoera eta ikerketaren helburua

80ko hamarkadatik aurrera itzulpen ikasketen alorrean, hainbat ikerlarik nabarmendu dute testu itzuliek ezaugarri bereizgarriak dituztela, itzulpen prozesuaren ondorioz. Ezaugarri horiek corpusetan oinarritutako azterketa kontrastiboen bidez deskribatu dira hainbat ikerlanetan (Hu, 2015).

Berriki, antzeko metodoak erabilita, lehen urratsak eman dira itzulpen automatikoen nolakotasuna aztertzeko. Izan ere, itzulpen automatikoek ere ezaugarri bereizgarriak dituztela egiaztatu da, testu originalekin ez ezik giza itzulpenekin ere alderatuta. Bizzoni eta kideek (2020) giza itzulpenen, giza interpretazioen eta itzulpen automatikoen aldeak aztertu zituzten, eta ondorioztatu zuten hirurek antzekotasunak badituzte ere, itzulpen automatikoak beste testuetatik bereiz daitezkeela.

Nolanahi ere, ebidentziak iradokitzen du giza itzulpenetan ez bezala, ezaugarri horiek ez direla itzulpen prozesuaren ondorio hutsa, eta algoritmoen gabezietatik eratorritakoak izan daitezkeela. Vanmassenhove eta kideek (2019) hiru itzultzaile automatiko desberdin entrenatu zituzten, bi sistema neuronal eta estatistiko bat, eta sistemek sortutako itzulpen automatikoak entrenamendu corpuseko giza itzulpenekin alderatu zituzten. Zehazki, aniztasun lexikoa konparatu zuten, metrika automatikoak erabilita. Emaitzek agerian utzi zuten algoritmoek entrenamendu corpuseko alborapenak erreproduzitzeaz gain, anplifikatu ere egiten dituztela, hots, hizkuntza egitura erabilienak hobesten dituztela, gutxi erabilitako egituren kalterako. Geroagoko lan batean, aniztasun morfologikoa eta sinonimoen erabilera ere aztertu zituzten, eta berriz ere aniztasun galera bat zegoela behatu zuten (Vanmassenhove et al, 2021).

Antzeko ikerketa batean, Shaitarova eta kideek (2024) itzultzaile komertzialen eta giza itzulpenen aniztasun lexiko, morfologiko eta sintaktikoa alderatu zituzten, eta behatu zuten kasu guztietan giza itzulpenak zirela aniztasun handienekoak, nahiz eta kasu batzuetan itzultzaile automatikoek gertuko emaitzak lortu zituzten.

Lan zehatzago batean, Gamallo eta kideek (2020) ingelesetik gaztelaniarako ahots pasiboaren itzulera aztertu zuten, eta ondorioztatu zuten gaztelaniak hiru egitura eskaintzen baditu ere, itzultzaile automatikoek ingelesetik gertuen dagoen egitura hobesten dutela orokorrean.

Ezaugarri horien zuzenketari dagokionez, badirudi giza berrikuspena ez dela nahikoa euren presentzia ezabatzeko. Toral eta kideek (2019) itzultzaile profesionalek posteditatutako testuak aztertu zituzten eta ondorioztatu zuten haietan oraindik ere ezaugarri bereizgarri horiek ageri direla.

Iraunkortasunerako joera horrek itzulpen automatikoen ezaugarriak sakon aztertzeko beharra azpimarratzen du, bereziki euskara bezalako hizkuntza gutxituetan. Hala ere, aurretik aipatutako lanak hizkuntza nagusituetan zentratu dira. Halaber, metodologia aldetik giza itzulpenen eta itzulpen automatikoen alderaketan zentratu dira, testu originalak alde batera utzita. Nolanahi ere, itzulpen automatikoak giza hizkuntzan izan dezakeen inpaktua aztertzeko, funtsezkoa da sorkuntza originalen eta itzulpen automatikoen arteko erlazioa ere ulertzea.

Gure lanak ezagutza-hutsune horiek betetzeko lehen urrats bat ematea du helburu. Horregatik, metodologia aldetik, aurreko lanekin alderatuta, gure lanerako abiapuntu desberdin bat aukeratu eta zuzenean testu originalekin lan egin dugu. Horrez gain, aurreko lanek ez bezala, gure ikerlana hizkuntza gutxitu baten testuinguruan zentratu dugu, euskararenean hain zuzen ere. Konkretuki sinonimoen erabilera aztertu dugu, dibertsitate lexiko-semantikoari zuzenean eragiten dion faktorea baita. Oro har, sinonimoen azterketari dagokionez gure lanak bi

helburu zehatz izan ditu: (1) ea sinonimoen erabileran euskaran ere aniztasun galerarik ba ote dagoen aztertzea eta (2) aniztasun galerarik egotekotan, galeraren adibide konkretuak identifikatzea.

3. Ikerketaren muina

3.1 EITB Corpusa

Aurreko atalean azaldu den bezala, aurretik egindako ikerketak itzulpen automatikoen eta giza itzulpenen arteko desberdintasunetan zentratu dira. Horren ondorioz, corpus paraleloak erabili dituzte azterketarako, hau da, sorburu-testuez eta euren itzulpenez osatutako corpus lerrokatuak erabili dituzte (Baker, 1995). Dena dela, abiapuntu hori desegokia da gure helburuei heltzeko, testu paraleloen bidez ezin baititugu zuzenean testu originalak aztertu. Oztopo hori gainditzeko, corpus konparagarri elebidun bat erabiltzea erabaki dugu. Mota horretako corpusak sortzeko, irizpide zehatz batzuei jarraituta, ezaugarri komunak (gaia, testu generoa…) dituzten bi hizkuntza desberdinetako eta independenteki sortutako testuak biltzen dira (Baker, 1995).

Zehazki EITB corpusaz (Etchegoyhen et al., 2020) baliatu gara, zeinak 2009 eta 2019 artean EITBko webgunean argitaratutako euskarazko eta gaztelaniazko berriak biltzen dituen. Bi hizkuntzetako albisteak modu independentean sortu ziren, baina gertaera berberen ingurukoak dira. Corpus hau itzultzaile automatikoen garapenerako sortu zenez, testuak prozesatzerakoan euskara eta gaztelaniazko testuak esaldika segmentatu ziren, eta ondoren edukiaren aldetik elkarren baliokideak izan zitezkeen esaldiak identifikatu eta lerrokatu ziren. Baliokide argirik gabeko segmentuak alde batera utzi ziren, eta baita ezarritako gutxieneko antzekotasun atalasera iritsi ez ziren segmentuak ere. Prozesaketa hori abantaila bat da gure lanerako, segmentuek antzeko informazioa transmititzen duten heinean, corpusa segmentu mailan ere konparagarria delako. Gainera, corpusaren tamaina egokia da gure aurkikuntzen fidagarritasuna sostengatzeko. Izan ere, 1. taulan adierazten den bezala, corpusak 600.000 esaldi baino gehiago ditu.

1. taula. Corpusaren azterketa kuantitatiboa

	Esaldiak	Hitz kopuru gordina	Hitz ezberdin kopurua
Gaztelaniazko originala	637.182	11.690.995	256.547
Euskarazko originala	637.182	8.506.695	225.593
Euskarazko itzulpen automatikoa	637.182	8.080.937	231.534

EITB corpuseko gaztelaniazko testuetatik abiatuta, albisteen itzulpen automatikoak lortu genituen ELIA itzultzailea erabilita. Alborapenik gabeko emaitzak lortzeko, ezinbestekoa zen EITB corpusa itzultzaile automatikoaren entrenamendu datuetatik kanpo egotea. Gure lanak baldintza hori betetzen du; izan ere, ELIAren entrenamendu-corpusak ez du EITB corpusa barne hartzen.

Behin itzulita, testua ixaKat[1] tresna erabilita prozesatu genuen, corpusa automatikoki aztertu ahal izateko. Konkretuki, tokenizatu, lematizatu eta kategoria gramatikalen mailan etiketatu genuen.

3.2 Metodologia

Vanmassenhove eta kideek (2021) itzulpen automatikoen eta giza itzulpenen arteko sinonimoen erabilera alderatu zuten metrika automatikoak erabilita. Horretarako, izen, aditz, eta adjektibo guztiak sorburu testutik erauzi eta hiztegi elebidun bat erabilita euren itzulpen posibleekin parekatu zituzten. Ondoren, jatorri hizkuntzako hitzen eta bere itzulpen posibleen frekuentziak kalkulatu zituzten, bai sistema entrenatzeko corpusean bai itzultzaile automatikoek sortutako testuetan. Sinonimo multzo bakoitzeko aniztasuna neurtzeko 3 metrika sortu zituzten:

[1] Prozesaketa honen emaitzak ez dira % 100ean fidagarriak. Hala ere, Eustaggerrek, ixaKaten analizatzaileak, kategoria gramatikalen etiketatzean % 95.17ko zehaztasuna du, eta % 91.89koa etiketatze morfologikoan (Otegi et al, 2016). Hortaz, akats batzuk egon daitezkeen arren, orokorrean emaitzak fidagarriak dira.

itzulpen maiztasun primarioa (PTF), sinonimoen banaketa uniformerako kosinu distantzia (CDU), eta **s**inonimoen *Type-Token Ratio*-a (SynTTR).

Euren metodologia jarraituz, guk ere sinonimoen erabilera aztertzeko hiru metrika horiek erabili ditugu. Dena den, sinonimo multzoak era desberdin batean osatu ditugu. Alde batetik, gaztelaniazko sorburu testua erabili beharrean, sinonimo multzoak zuzenean euskarazko bi testuetan oinarrituta sortu ditugu Euskarazko Wordnet-en laguntzaz (Pociello et al., 2011). Bestetik, Vanmassenhove eta kideen (2021) lanean itzulpen posible guztiak hartzen dira aintzat, hitzaren kategoria edo zentzua kontuan hartu gabe, baina gure lanean sinonimoen multzoa hitz kategoriaren arabera mugatuta dago.

Moldaketa horiek barneratzea ezinbestekoa zen gure ikerlanerako, gure corpusa eta Vanmassenhove eta kideek (2021) erabiltzen dutena desberdinak direlako. Euren corpus paraleloan ez bezala, non sorburu testuak itzulpen automatikoekin lotzen diren, gure corpus konparagarrian soilik itzulpen automatikoek dute lotura estua sorburu testuarekin, testu originalak ez baitira bertatik eratorri. Hori dela eta, ezin ditugu zuzenean gaztelaniazko ordainak testuetan ageri diren sinonimoekin erlazionatu. Alderaketa esanguratsuak lortzeko, elkarrekin truka daitezkeen eta antzeko testuinguruetan ager daitezkeen sinonimoak lehenetsi ditugu. 2. taulan Vanmassenhove eta kideek (2021) erabilitako estrategia gure estrategiarekin alderatzen da.

2. taula. Sinonimoak multzokatzeko metodoen alderaketa.

Vanmassenhove et al. (2021)	Hitza: “look” Sinonimoen multzoa: {“mirar”, “esperar”, “buscar”, “parecer”, “dar”, “vistazo”, “aspecto”, “ojeada”, “mirada”}
Gure metodoa	Hitza: “desberdindu » Sinonimoen multzoa: {“desberdindu”, “ezberdindu”, “bereizi}

Estrategia hori erabilita, 1.462 izen sinonimo multzo eta 184 aditz multzo sortu ditugu. Hasieran, Vanmassenhove eta kideen (2021) antzera, adjektiboen multzoak ere sortzeko asmoa genuen, baina euskarazko WordNet-eko adjektiboen sarrera eskasen eta gure multzokatze kriterio zorrotzen ondorioz, bederatzi multzo baino ezin izan ditugu lortu. Horregatik, adjektiboen azterketa lan honetatik kanpo utzi dugu.

Sinonimo multzoak eraiki eta euren frekuentziak eskuratu ondoren, aurretik aipatutako hiru metrikak erabilita (PTF, CDU eta SynTTR) ebaluatu ditugu.

PTF metrikak gehien erabilitako itzulpenak beste aukeren gainetik duen nagusitasuna neurtzen du. Metrika horrek iradokitzen du aukera bat besteen gainetik sistematikoki aukeratzean, aniztasuna galtzen dela. PTF balio txikiagoek adierazten du aukera erabiliena ez dela besteengandik hainbeste gailentzen.

CDUk sinonimo distribuzioen uniformitatea neurtzen du. Horretarako distribuzioaren eta distribuzio uniforme baten arteko kosinuaren distantzia kalkulatzen du. CDU balio txikiagoek sinonimoen erabilera orekatuagoa adierazten dute.

SynTTR *type-token ratio* metrikatik eratorritako metrika bat da. Kasu honetan, *type*ak sinonimo desberdinen kopuruari dagozkio, eta *token* kopurua euren maiztasunari. Itzultzaile automatikoak beste sinonimo posibleak borratzen dituen kasuen identifikaziorako baliagarria da. Ohiko TTR-an bezala, balio altuagoek aniztasun handiagoa adierazten dute. Ikusi 3. taulan metriken kalkuluen adibideak.

3. Taula. PTF, CDU eta SynTTR metriken kalkuluaren adibideak.

Sinonimo-distribuzio baten adibidea
{“desberdindu”: 30, “ezberdindu”: 69, “bereizi”: 281}
PTF metrikaren kalkuluaren adibidea
$\frac{281}{30+69+281} = 0.739$
CDUren kalkuluaren adibidea
Bektore uniforme baten kalkulua: $[\frac{30+69+281}{3}, \frac{30+69+281}{3}, \frac{30+69+281}{3}] = [126.67, 126.67, 126.67]$

CDUren kalkulua: Spicy.spatial.distance.cosine([30, 69, 281], [126.67, 126.67, 126.67]) = 0.245
SynTTRren kalkulua
$\frac{3}{30 + 69 + 281} = 0.0078$

3.3 Emaitzak

Metrikak indibidualki kalkulatu ditugu sinonimo multzo bakoitzerako, eta guztizko emaitzak batez bestekoak kalkulatuz lortu dira. 4. taulan amaierako emaitzak aurkezten ditugu izen eta aditzen multzoetarako. Metrika hauentzako balioa 0 eta 1 artekoa da.

4. taula. Sinonimia metriketan lortutako emaitzak.

		Euskarazko originala	Euskarazko itzulpen automatikoa
Izenak	PTF ↓	**0.82**	0.84
	CDU ↓	**0.23**	0.24
	SynTTR ↑	**0.12**	0.10
Aditzak	PTF ↓	**0.78**	0.80
	CDU ↓	**0.27**	0.28
	SynTTR ↑	**0.49**	0.38

Emaitzek adierazten dute testu originalek aniztasun handiagoa dutela hiru metriketan, bai aditzen bai izenen sinonimo multzoentzako. Dena den, aldeak zeharo apalak dira. Garrantzitsua da azpimarratzea gure multzokatze kriterio zorrotzaren ondorioz soilik hitz monosemikoak aztertu ditugula, eta, hortaz, soilik sinonimiaren frakzio bat baino ez dugula aztertu. Hitz polisemikoak barneratuz gero, emaitza zabalagoak eskuratuko genituzke.

Sinonimo erabilienei erreparatzean, behatu dugu kasu gehienetan itzulpen automatikoan eta testu originalean gehien erabilitako sinonimoa berdina dela. Konkretuki, sinonimo erabilienak bat egiten du giza hobespenekin izen multzoen % 85ean eta aditz multzoen % 83,6an. Dena dela, itzultzaile automatikoan leheneste hori nabarmenagoa da, lehenetsitako sinonimoaren presentzia % 3,76 handitzen baita.

Behaketan sakontzeko, gutxienez 50 aldiz agertzen diren sinonimo multzoak zehaztasun handiagoz behatu ditugu. 5. taulan izen eta aditz multzoen adibideak bildu ditugu. Ezkerreko zutabean sinonimoak zerrendatu ditugu eta eskuinera dauden zutabean euren frekuentzia jaso dugu. Azterketak agerian uzten du itzulpen automatikoak ezohiko edo domeinu espezifikoetako hitzetan alborapena izateaz gain, eguneroko hitzetan ere ageri duela. Adibidez, “guraize” eta “artazi” hitzen artean, nabari da itzultzaile automatikoak lehenengoa nahiago duela.

5. taula. Sinonimo multzoen adibideak.

	Sinonimo multzoak?	EU-O	EU-MT
Izenak	[‘guraize’, ‘artazi’]	[19 25]	[46 1]
	[‘margo’, ‘pintura’]	[81 135]	[2 279]
	[‘galtza’, ‘praka’]	[38 32]	[60 3]
	['margolan', 'pintura', 'koadro']	[108 135 94]	[2 279 191]
	[‘okela’, ‘haragi’]	[43 120]	[6 162]
	['gaztelania', 'espainol', 'espainiera', 'gaztelera']	[550 8 13 218]	[751 3 6 1]
Aditzak	[‘muturtu’, ‘sumindu’, ‘asaldatu’, ‘haserretu’]	[2 94 42 79]	[0 21 15 168]
	[‘espetxeratu’, ‘kartzelatu’]	[240 114]	[140 0]
	[‘desberdindu’, ‘ezberdindu’, ‘bereizi’]	[30 69 181]	[21 2 574]
	['gerturatu', 'inguratu', 'hurbildu', 'ondoratu', 'alboratu', 'alderatu', 'hurreratu']	[493 240 1.238 4 48 1.168 20]	[14 294 1.466 0 12 838 0]
	['prebenitu', 'ekidin', 'saihestu', 'eragotzi']	[54 543 1.069 307]	[103 33 1.395 608]

4. Ondorioak

Lan honetan EITB corpusa automatikoki itzuli dugu, eta gizakiek sortutako euskarazko albiste originalak eta itzulpen automatikoak alderatu ditugu. Konkretuki, hizkuntzaren aniztasunari zuzenean eragiten dion fenomeno batean zentratu da: sinonimian. Sinonimoen erabilera alderatzeko, metodo automatikoaz baliatu gara, lehenik, sinonimo posibleak testuetatik erauzteko eta euren frekuentzia kalkulatzeko, eta ondoren, sinonimo-distribuzioen hizkuntza aniztasuna neurtzeko.

Izen eta aditz sinonimo multzoen aniztasuna aztertzeko hiru metrika erabili ditugu: PTF, CDU eta SynTTR. Metrika horietan eskuratutako emaitzek iradokitzen dute orokorrean sinonimoen aldetik bi testuek antzeko aniztasuna dutela, nahiz eta oro har testu originaletako sinonimo distribuzioak orekatuagoak diren, eta, hortaz, aniztasun handiagoa duten. Halaber, gizakien eta itzultzaile automatikoaren hobespenak gehienetan bat datozela antzeman dugu, hots, gizakiek gehien erabilitako aukerak itzultzaile automatikoak ere gehien erabiltzen dituenak direla. Azkenik, adibide konkretuak aztertu, eta itzultzaile automatikoaren alborapen zehatzak identifikatu ditugu. Horri lotuta termino oso ohiko batzuetan ere alborapena ageri dela behatu dugu.

5. Etorkizunerako planteatutako norabidea

Ikerlan honetan egindako ikerketa hizkuntzaren alderdi konkretu batean zentratzen da, sinonimoen azterketan, alegia. Itzulpen automatikoen ezaugarri bereizgarriak hobeto ulertzeko mota honetako azterketak egin beharko lirateke eta hizkuntzaren beste alderdi batzuk aztertu, hala nola sintaxia, morfologia eta beste alderdi lexiko-semantiko batzuk. Halaber, hemen aurkeztutako metodologiak sinonimoa modu hertsian aztertzen du, hitz kategoria bereko zentzu bakarreko sinonimoak baino ez baitira aztertzen. Etorkizunean, interesgarria litzateke sinonimia modu zabalago batean aztertzea, hitz kategoria aldaketak eta hitz polisemikoak barne hartzeko.

6. Erreferentziak

Aranberri, N., & Iñurrieta, U. (2024). When minoritized languages encounter MT: perceptions and expectations of the Basque community. *The Journal of Specialised Translation*, (41), 179-205.

Baker, M. (1995). Corpora in translation studies. *Target. International Journal of Translation Studies, 7*(2), 230–230. https://doi.org/10.1075/target.7.2.03bak

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins.

Castilho, S., Resende, N., Mitkov, R. (2019). What influences the features of posteditese? A preliminary study. In Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019) (pp. 19–27). Incoma Ltd. https://doi.org/10.26615/issn.2683-0078.2019 003

Bizzoni, Y., Juzek, T.S., España-Bonet, C., Dutta Chowdhury, K., van Genabith, J. & Teich, E. (2020). How human is machine translationese? Comparing human and machine translations of text and speech. In M. Federico, A. Waibel, K. Knight, S. Nakamura, H. Ney, J. Niehues, S. Stüker, D. Wu, J. Mariani, F. Yvon (Eds.), *Proceedings of the 17th International Conference on Spoken Language Translation* (pp. 280–290). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.iwslt-1.34

Etchegoyhen, T., Harritxu, G. (2020). Handle with care: A case study in comparable corpora exploitation for neural machine translation. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3799–3807). European Language Resources Association. https://aclanthology.org/2020.lrec-1.469

Gamallo, P., Labaka, G. (2021). Using dependency-based contextualization for transferring passive constructions from english to spanish. *Procesamiento del lenguaje natural, 66*, 53– 64. https://doi.org/10.26342/2021-66-4

Hernáez, I., Navas, E., Odriozola, I., Sarasola, K., Diaz de Ilarraza, A., Leturia, I., Diaz de Lezana, A., Oihartzabal, B., Salaberria, J. (2012). *The Basque Language in the Digital Age / Euskara Aro Digitalean*. Springer.

Hu, K. (2015). Introducing corpus-based translation studies. In *New frontiers in translation studies*. https://doi.org/10.1007/978-3-662-48218-6

Otegi, A., Ezeiza N., Goenaga, I., Labaka, G. (2016). A Modular Chain of NLP Tools for Basque. In P. Sojka, A. Horák, I., Kopeček (Eds.), *Proceedings of the 19th International Conference of Text, Speech, and Dialogue, SD 2016, Brno, Czech Republic, Lecture Notes in Computer Science* (pp. 93-100). Springer International Publishing. https://doi.org/10.1007/978-3-319-45510-5_11

Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and Construction of the Basque WordNet. *Language resources and evaluation, 45*(2), 121–142.

Shaitarova, A., Göhring, A., Volk, M. (2023). Machine vs. human: Exploring syntax and lexicon in German translations, with a spotlight on anglicisms. In T. Alumäe, M. Fishel (Eds.), *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 215–227). University of Tartu Library. https://aclanthology.org/2023.nodalida-1.22

Toral, A. (2019). Post-editese: an exacerbated translationese. In M. Forcada, A. Way, B. Haddow, R. Sennrich (Eds.), *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 273–281). European Association for Machine Translation. https://aclanthology.org/W19-6627

Vanmassenhove, E., Shterionov, D., Way, A. (2019). Lost in translation: Loss and decay of linguistic richness in machine translation. In M. Forcada, A. Way, B. Haddow, R. Sennrichar (Eds.), *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 222–232). Association for Computational Linguistics .URL: https://aclanthology.org/W19-6622

Vanmassenhove, E., Shterionov, D., Gwilliam, M. (2021). Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2203-2213). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.188

Zabaleta, Josu. (2019). Itzulpengintza eta euskararen batasuna eta normalizazioa. Mende erdiko historiaren berrikusketa eta gogoeta batzuk. *Senez, 50*, 85–10

7. Eskerrak eta oharrak

Lan hau Eusko Jaurlaritzako Hezkuntza, Unibertsitate eta Ikerketa Sailak (PRE-2024-1-0147) finantziatu du, eta MCIN/AEI/10.13039/501100011033 eta FEDERrek (Europa egiteko modu bat) finantziatutako PID2021-123988OB-C31 proiektuaren barnean egin da.