Universidad del País Vasco Euskal Herriko Unibertsitatea

Hiztegi elebidunak automatikoki sortzea ahalbidetzen duten pibotaje tekniken azterketa eta konparaketa

Egilea: Iker Manterola Isasa
Tutorea: Kepa Sarasola Gabiola

hap

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua lortzeko bukaerako proiektua

2012eko iraila

Sailak: Lengoaia eta Sistema Informatikoak, Konputagailuen Arkitektura eta Teknologia, Konputazio Zientziak eta Adimen Artifiziala, Euskal Filologia, Elektronika eta Telekomunikazioak.

Laburpena

Hiztegi elebidunak oinarrizko baliabideak dira itzulpengintzan, hizkuntzen ikasketan edota hizkuntza naturalen prozesamenduaren arloko hainbat atazatan. Hala ere, baliabide gutxien duten hizkuntzen kasuan oso zaila da mota honetako baliabideak eskuratzea. Testuinguru honetan, hiztegi elebidunak automatikoki sortzea ahalbidetzen duten tresna edota teknikak erabiltzea oso lagungarria izan daiteke. Lan honetan hiztegi elebidunak automatikoki sortzeko erabili daitezkeen teknikak aztertu eta konparatzen dira, eta baita horien bidez lortzen diren hiztegiak ere. Azterketa horiek baliabide gutxien duten hizkuntzen testuinguruan kokatzen dira.

Abstract

Bilingual dictionaries are key resources in several fields such as translation, language learning or various NLP tasks. However, it is very difficult to obtain this kind of resources in the case of less resourced languages. In this context, the use of techniques and tools which allow to create bilingual dictionaries automatically can be very useful. This work analyzes and compares these techniques, and also the dictionaries obtained with them. These analyses are performed in the context of less resourced languages.

Eskerrak,

Elhuyar Fundazioko I+G saileko lankide diren *Iñaki San Vicente* eta *Xabier Saralegi*-ri, lan honetan aurkezten diren edukiak biltzen dituen *Pibolex* proiektuan egindako lanagatik eta emandako laguntzagatik.

Gaien aurkibidea

Irudien zerrenda

Taulen zerrenda

1 Sarrera

Hiztegi elebidunek hainbat eta hainbat erabilpen esparrutan duten erabilgarritasuna ezin da zalantzan jarri. Hizkuntza baten ikasketa-prozesuan, itzulpen lanetan edo edozein motako testu baten ulertze-prozesuan oinarrizko baliabidea dira. Antzeko zerbait esan daiteke hizkuntza naturalen prozesamenduaren arloan betetzen duten funtzioaz ere: ia edozein atazaren beharrezko baliabideen artean hiztegi elebidunak daude (itzulpen automatikoan, informazioaren berreskurapen eta erauzketan, ...). Hala ere, hiztegi elebidun horien sorkuntza oso garestia eta neketsua da oraindik ere, eta ondorioz edozein hizkuntza bikoteren arteko hiztegi elebidunak lortzea ia ezinezko lan bihurtzen da. Are gehiago landu nahi diren hizkuntzen artean baliabide gutxiko hizkuntza bat agertzen denean.

Aipatu berri den testuinguruan, hiztegi elebidunak automatikoki edo behintzat erdiautomatikoki lortzeko balio dezaketen tresnak edota prozesuak sortzea beharrezko bihurtzen da. Argitu beharra dago hiztegi elebidunak aipatzen ditugunean sarrera-itzulpena egitura duten hiztegiei egingo zaiela erreferentzia: lan honetan aipatzen diren hiztegietan ez da definiziorik erabiltzen, hitz bakarreko itzulpenak baizik.

Azken urteetan Internet sareak izan duen hazkundea kontuan hartuta, bertako baliabideen ustiapenetik hiztegi elebidunak sortzea ideia erakargarria bihurtu da. *Wikipedia* edota *Wiktionary* bezalako baliabideak behar hori asetzeko hautagai bikainak dirudite. Zoritxarrez, baliabide horietan hainbat hizkuntzek duten presentzia benetan handia izan arren (ingelesa, txinera, ...), beste hainbat hizkuntzena ez da maila erabilgarri batera heltzen (euskara, txekiera, ...). Beraz, eta oraingoz behintzat, sareko baliabide horiek ez dira aukera egokiena.

Baliabide falta honen aurrean, azken urteetan pibotaje tekniken erabilpena indarra hartzen joan da. Teknika horien oinarria sinplea bezain indartsua da: erlaziorik ez duten bi hizkuntza erlazionatzeko nahikoa da hirugarren hizkuntza bat zubi bezala erabiltzea; izan ere, kontsidera daiteke hiztegi sarreren itzulpenek hizkuntza arteko propietate trantsitiboa mantentzen dutela. Adibide erraz bat jartzearren, euskara eta gaztelera barneratzen dituen hiztegi elebidun bat sortu nahiko bagenu, ingelesa erabil genezake zubi bezala. Horrela, euskarazko *"zakur"* sarreraren itzulpena ingelesezko *"dog"* hitza bada, eta horren gaztelerazko itzulpena *"perro"*, esan genezake euskarazko *"zakur"* sarreraren gaztelerazko itzulpena *"perro"* dela. Ideia erraza izan arren, teknikak hainbat arazo nabarmen ditu. Lan honetan pibotaje tekniken azterketa sakon bat egingo da, baliabide gutxien duten hizkuntzen testuinguruan kokatuz.

Bukatzeko, aipatu lan honetan aurkeztuko diren lan eta emaitzetan oinarrituta bi artukulu publikatu direla bi kongresu ezberdinetan: Lehena *EMNLP 2011 (Conference on Empirical Methods in Natural Language Processing)* kongresuan, *"Analizing Methods for Improving Precision of Pivot Based Bilingual Dictionaries"* izenburupean, eta bigarrena *LREC 2012 (In Proceedings of the 8th international conference on Language Resources and Evaluation)* kongresuan, *"Building a Basque-Chinese Dictionary by using English as a Pivot"* izenburupean.

2 Aurrekariak

Kapitulu honetan zehar lan honetan oinarritzat hartzen diren lan eta teknika guztiak azalduko dira. Horretarako, lehendabizi pibotaje tekniken oinarrizko kontzeptuak aurkeztuko dira.

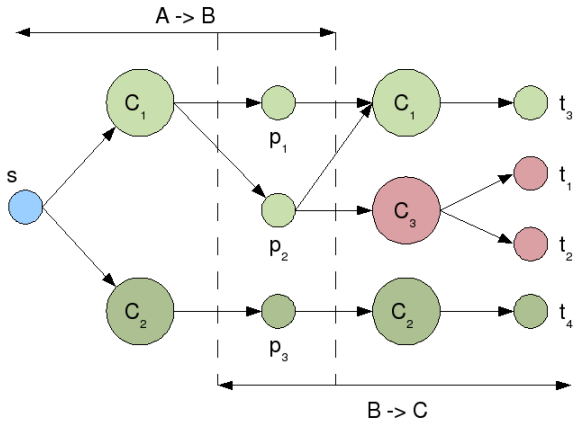
2.1 Pibotaje teknikak

Bi hizkuntza konkreturen arteko lexikoa hiztegi elebidun batean lortu ezin denean, lexiko hori lortzeko aukera bat hirugarren hizkuntza bat zubi moduan erabiltzea izaten da. Adibidez, itzultzaile batek domeinu zehatz bateko sarrera baten itzulpena beste hizkuntza batean zuzenean lortu ezin duen kasuetan, hirugarren hizkuntza batean eskuz bilatu ohi du horrela jatorrizko sarreraren esanahia ulertu ahal izateko. Prozedura hori da hiztegi elebidunak automatikoki sortzea ahalbidetzen duten pibote tekniken oinarrian dagoen ideia, bi hiztegi elebidun gurutzatuz (*A-B* eta *B-C*) hirugarren hiztegi elebidun bat lortzea alegia *(A-C)*. Azken finean, hiztegi sarreren itzulpenek hizkuntza arteko propietate trantsitiboa mantentzen dutela suposatzen da, eta beraz jatorrizko A hizkuntza bateko a sarrera baten itzulpena B hizkuntzan b bada, eta aldi berean b itzulpenaren C hizkuntzako itzulpena c bada, orduan suposa daiteke C hizkuntzako c A hizkuntzako a sarreraren itzulpena dela:

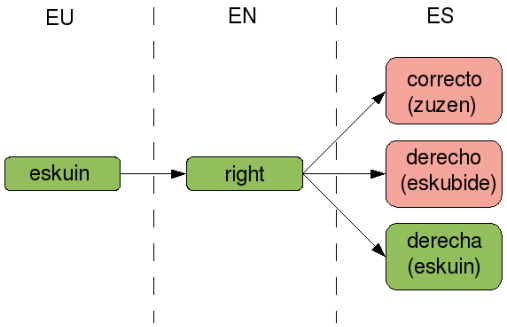
$$itzulpenaA\text{-}B(a)=b \wedge itzulpenaB\text{-}C(b)=c \rightarrow itzulpenaA\text{-}C(a)=c$$

Sinplifikazio hori ez da guztiz zuzena; izan ere, hitz batek eduki ditzakeen adiera guztiak ez ditu kontuan hartzen. Demagun 2.1 irudian ageri den A hizkuntzako s hitzak bi adiera posible dituela (C_1 eta C_2). C hizkuntzan lortutako itzulpenei begiratuz gero (t_1, t_2, t_3, t_4), C_1 (t_3) eta C_2 (t_4) adierez gain C_3 adiera duten itzulpenak ere lortzen direla ikus daiteke (t_1, t_2). C_3 adiera s hitzaren adiera posibleen artean ez dagoenez, t_1 eta t_2 itzulpenak ez dira zuzenak. Jatorrizko hitzei ez dagozkien adiera berri horiek prozesuan parte hartzen duten hitzen polisemiaren ondorioz sortzen dira (batez ere zubi hizkuntzako polisemiaren eraginez). 2.1 irudiko adibidean zubi hizkuntzako p_2 hitza da polisemiaren ondorioz adiera berri bat gehitzen duena.

Azter dezagun benetako adibide bat (2.2 irudia): euskarazko *"eskuin"* sarrera, helburu hizkuntza gaztelera izanik eta zubi bezala ingelesa erabiliz. Euskarazko *"eskuin"* hitzaren ingelesezko itzulpenak aztertzen baditugu, *"right"* hitza aurkituko dugu, eta honen adierak aztertuz konturatuko gara ingelesez gutxienez hiru adiera ezberdin dituen hitz baten aurrean gaudela (euskaraz duena eta beste bi): *"eskuin"* adiera, *"zuzen"* adiera eta *"eskubide"* adiera. Horrela, ingelesezko hitz honen gaztelerazko itzulpenen artean *"derecha"*, *"correcto"* eta *"derecho"* (*"eskubide"*) hitzak agertuko dira. Pibotajearen ideia aplikatuz gero, euskarazko *"eskuin"* hitzaren itzulpenak *"derecha"*, *"correcto"* eta *"derecho"* direla ondorioztatuko genuke, eta kasu honetan nabaria da ondorio hori ez dela zuzena. Beraz, prozesuan parte hartzen duten hizkuntzen polisemia mailak berebiziko eragina du pibotaje-prozesuetan (batez ere zubi moduan jokatzen duen hizkuntzaren polisemia mailak), honen ondorioz itzulpen okerrak sortzeko arriskua handia baita.



Irudia 2.1: Polisemiaren eragina pibotaje bidezko hiztegien sorkuntza-prozesuan



Irudia 2.2: Polisemiaren eragina pibotaje-prozesuetan

2.2 Polisemiaren eragina murrizten

Pibotaje bidezko hiztegi-sorkuntzan polisemiak duen eragina murrizteko hainbat eta hainbat irtenbide proposatu izan dira azken urteetan. Horretarako, autore batzuek gurutzaketan parte hartzen duten hiztegietan aurki daitekeen informazioan bakarrik oinarritzen diren bitartean, beste hainbatek hiztegiez gain bestelako kanpo baliabide batzuk erabiltzea proposatzen dute. Jarraian azken urteetan proposatutako teknikak aurkeztuko dira.

2.2.1 Hiztegietako informazioa bakarrik erabiliz

Teknika edo prozedura hauek gurutzatzen diren hiztegietan lor daiteken informazioaren erabilpenean soilik oinarritzen dira. Hori dela eta, teknika hauen erabilpena prozesuan parte hartzen duten hiztegien ezaugarrien menpe dago, jakina baita hiztegi guztien edukiek ez dutela mota bereko informazioa eskaintzen. Jarraian hiztegietatik lor daitekeen informazioan oinarritzen diren teknikak sakonago aztertuko dira.

2.2.1.1 *Inverse Consultation* (Alderantzizko kontsulta)

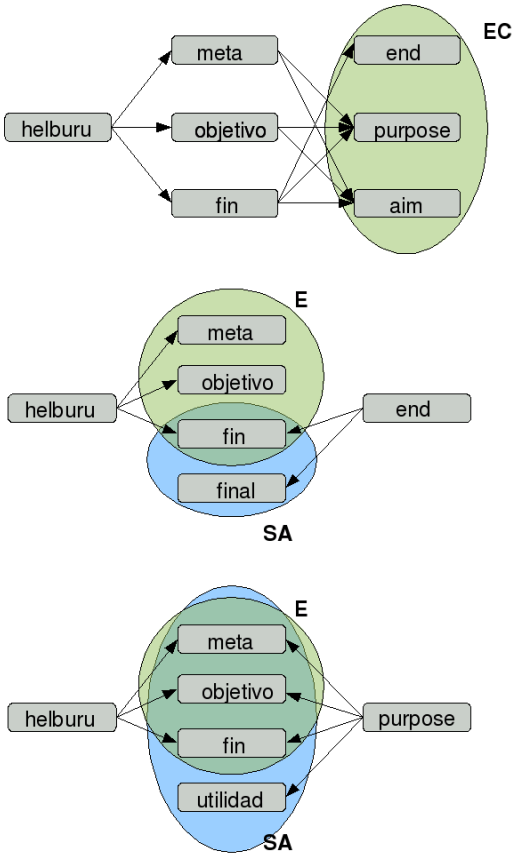
Tanaka eta Umemura (1994) lanean, gurutzaketan parte hartzen duten hiztegien egituretan oinarritzea proposatzen dute, polisemiaren ondorioz sorturiko itzulpen okerrak baztertzeko. Horretarako, lehenik eta behin sarrera baten itzulpen posible guztiak eskuratzen dituzte pibote edo zubi hizkuntzan *(E – baliokidetzak)*, eta horietariko bakoitzeko helburu hizkuntzan dituzten itzulpen posible guztiak *(EC – baliokide hautagaiak)*. Jarraian, hautagai bakoitza jatorrizko hiztegiak duen kontrako norabidean bilatzen da *(SA – hautatze area)*. *SA* eta *E* multzoek komunean duten itzulpen kopuruak sarrerak bakoitza eta bere itzulpen posibleen arteko antzekotasun semantikoa estimatzeko balioko du. Zehazki, multzo horien ebakidurak elementu bat baino gehiago duen kasuetan itzulpen hautagaiak onargarriak direla kontsideratuko da, eta jakina, geroz eta elementu gehiago izan komunean orduan eta itzulpen sendoagoa dela kontsideratuko da. Teknika honek *Inverse Consultation (IC)* izena jasotzen du (alderantzizko kontsulta). Aurkako norabidean kontsulta edo jauzi bakarra egiten bada, erabili den teknika IC_1 dela esaten da; aldiz, n kontsulta edo jauzi egiten badira aurkako norabidean IC_n dela esaten da. 2.3 irudian *"helburu"* hitzaren adibide bat ikus daiteke non IC_1 teknikaren arabera ingelesezko *"purpose"* eta *"aim"* itzulpenak zuzenak diren baina *"end"* ez. Tanaka eta Umemura (1994) lanean lorturiko emaitzak hiztegi argitaratuekin konparatzen dira eta emaitzen bidez hiztegi horiek aberastu daitezkeela ondorioztatzen da.

Tanaka eta Unemurak beraien lanean hiztegi harmonikoak erabili zituzten. Hiztegi bat harmonikoa dela esango dugu, sarrerak eta itzulpenak nodoak izanik, horiek lotzen dituzten arkuak bi norabidetakoak direnean edo, beste era batera esanda, hiztegia simetrikoa denean.

2.2.1.2 Kategoria gramatikal eta klase semantikoen erabilpena

Gurutzaketa arruntarekin edota *IC* teknikaren bidez osaturiko hiztegi berriak informazio lexikoan soilik oinarritzen dira, eta ondorioz kasu askotan horien doitasuna ez da espero bezain altua. Hori dela eta, zenbait autorek gurutzaketa-prozesuan zenbait hiztegitan aurki daitekeen informazio gehigarria erabiltzea proposatzen dute, izan ere, zenbait hiztegik sarrera eta itzulpenez gain horien informazio sakonagoa ematen dute; hala nola, kategoria gramatikalak edota klase semantikoak. Horrela, lorturiko emaitzen doitasuna hobetzeko bide bat gurutzatu diren hiztegietako sarreren kategoria gramatikalak erabiltzea da (*Elhuyar*[1] hiztegian adibidez *"perikito"* sarrera izen bezala markatua dago). Horrela, itzulpen bikote hautagai bakoitzeko sarrera eta itzulpenaren kategoria bat ez datorren kasuetan, informazio gehigarri horrek itzulpen bikotea okerra dela ondorioztatzen lagun dezake. Bond et al. (2001) lanean, gurutzaketa-prozesuan parte hartzen duten hiztegietako kategoria gramatikalen informazioa itzulpen bikoteen aukeraketa-prozesuan erabiltzen da, horrela itzulpen bikote ziurragoak lortzen direlarik. Kategoria gramatikalez gain zenbait hiztegitan topatu daitekeen beste informazio bat sarreren klase semantikoa da (aurreko adibidearekin ja-

[1] http://www.elhuyar.org/hizkuntza-zerbitzuak/EU/Hiztegi-kontsulta



Irudia 2.3: IC_1 teknikaren adibidea

rraituz, *Elhuyar* hiztegian adibidez *"perikito"* sarrera zoologia klasekoa bezala markatua dago). Berriro ere, sarrera baten informazio semantikoa gehitzeak itzulpen bikote hautagaien doitasuna hobetzen lagun dezake, itzulpen hautagai bateko sarrera eta itzulpena klase semantiko ezberdinekoak badira bikotea okerra izango baita. Horrela, Bond eta Ogura (2007) lanean, hiztegietako sarreren kategoria gramatikalen informazioa erabiltzeaz gain ontologia batetik lortutako klase semantikoen informazioa gehitzen da prozesura. Metodo horien bidez lortzen den doitasuna hobea izan arren, jakina da mota honetako baliabideak ez direla edozein hizkuntza konbinaziorako existitzen.

2.2.1.3 Hiztegietako definizioak

Hiztegi elebidunetan aurki daitekeen beste informazio erabilgarri bat sarreren testuzko definizioak dira. Definizio horiek hizkuntza ezezagun bateko sarrera bat hizkuntza ezagun bateko hitzak erabiliz deskribatzeko erabili ohi dira. Sjöbergh (2005) lanak definizio horiek informazio iturri bezala erabiltzea proposatzen du loturarik ez duten bi hizkuntza hirugarren hizkuntza baten bidez erlazionatzeko. Horretarako, Sjöbergh-en lanean definizioak

eskaintzen dituzten bi hiztegi erabiltzen dira: batak sarrera japoniarren ingelesezko definizioak ematen ditu eta besteak suedierazko sarreren ingelesezko definizioak. Horrela, japonierazko hitzen suedierazko itzulpenak lortu ahal izateko, ingelesezko definizioen antzekotasuna kalkulatzea proposatzen da. Antzekotasun hori neurtzeko, lehenik eta behin definizioetatik prozesuan laguntzen ez duten elementu guztiak ezabatzen dira (*stop-word*-ak, zenbakiak, ...) eta gainerako hitz guztiei pisu bat esleitzen die definizio zehatz bakoitzean hitz bakoitzak duen adierazgarritasuna adierazteko. Pisu hori neurtzeko *idf* neurria *(Inverse document frequency)* erabiltzen du. Neurri honen bidez hitz jakin batek dokumentu batean duen pisua neurtzen da dokumentu multzo batean duenarekiko. Horrela, japonierazko j hitzaren ingelesezko definizoan *(je)* eta suedierazko s hitzaren ingelesezko definizioan *(se)* errepikatzen diren hitzen pisuak erabiliz itzulpen bikote posible bakoitzari balio bat ematen dio. Balio horiekin rankingak osatzen dira, eta japonierazko sarrera bakoitzeko balio altuena duen suedierazko hitza itzulpen hoberena dela kontsideratzen da.

Bestalde, Sjöbergh (2005) lanak hizkuntza bateko sarrera zehatz bat beste hizkuntza bateko hitz bakarra erabiliz itzultzea ezinezkoa izan daitekeela kontuan hartzen du, eta oztopo hori gainditu ahal izateko sarrera bat hainbat hitz konbinatuz itzultzeko aukera ematen du. Era honetara, teknika honen bidez lortzen diren hiztegien estaldura handitzea lortzen du eta bestela hitz bakarra erabiliz itzultzea ezinezkoak izango liratekeen sarreren itzulpenak ematea lortzen du. Orokorrean teknika honen bidez lorturiko emaitzak nahiko onak dira, itzulpen zuzena existitzen den sarreratan balio altuena lortu duen itzulpena izaten delarik.

2.2.1.4 Hiztegien norabideak

Edozein hiztegik barneratzen duen informazio guztiaz gain, bestelako propietate garrantzitsu bat ere badu: norabidea. Edozein hiztegi elebidunek jatorrizko hizkuntza zehatz bateko sarreren itzulpenak helburu hizkuntza zehatz batean ematen ditu. Jatorri eta helburu hizkuntzek definitzen duten norabideak berebiziko eragina du hiztegien izaera eta portaeran (Shirai eta Yamamoto (2001); Shirai et al. (2001); Paik et al. (2004)). Adibidez, hizkuntza ezezagun bateko sarreren itzulpenak ematen dituen hiztegi bat, hizkuntza ezezagun horren lexiko guztia barneratzen saiatuko da. Hala ere, hizkuntza ezezaguneko sarrera zehatz batek hizkuntza ezagunerako itzulpenik ez badu sarrera hori ziurrenik ez da hiztegian agertuko. Bestalde, hitz bakarrarekin itzuli ezin diren hizkuntza ezezaguneko sarrerek azalpen luzeagoak izango dituzte (hau da, ez dira hitz bakarrera mugatuko), eta ondorioz ez dira erabilgarriak izango pibotaje teknika arruntekin, non ohikoena hitz bakarreko sarrera eta itzulpenekin lan egitea den.

Hiztegien norabideak kontuan hartuta, A eta B hizkuntzak C pibote hizkuntzaren bidez erlazionatzeko honako konbinaketak egin daitezke:

- **A→*C* + *C*→*B*:** Zubi edo pibote hizkuntza itzulpenak emateko erabiltzen da lehen hiztegian eta sarrerak emateko bigarrenean.
- ***A*→*C* + *B*→*C*:** Zubi hizkuntza itzulpenak emateko erabiltzen da bi hiztegietan.

- ***C→A + C→B*:** Zubi hizkuntza sarrerak emateko erabiltzen da bi hiztegietan.

Bestalde, hiztegien norabide ezberdinek sortzen dituzten desberdintasunen aurrean, Tanaka eta Umemura (1994) lanean hiztegi harmonizatu edo simetrikoak erabiltzea proposatu zuten *(A↔B)*. Hiztegi horien edukia berdina da bi norabideetan, eta horiek eraikitzeko $A \rightarrow B$ eta $B \rightarrow A$ hiztegien konbinaketa egiten da, emaitza bezala bi norabideetan betetzen diren itzulpen bikoteak bakarrik hartzen direlarik. Prozesu honek hiztegien estaldura txikiagotzen du, baina aldi berean erabiltzen diren hiztegien norabideek sortarazten dituzten diferentziak minimizatzen ditu.

Beraz, hiztegi bakoitzaren edukien izaeraz gain, horien norabideak ere gurutzaketaren bidez lortuko diren emaitzen ezaugarrietan eragin zuzena izango du.

2.2.2 Kanpo baliabideak erabiliz

Gurutzaketan parte hartzen duten hiztegietan aurki daitekeen informazio mota eta kopurua nahiko mugatua izaten da. Gainera, kasu askotan erabiltzen diren hiztegietan informazio hori ez da eskuragarri egongo, eta eskuragarri egonda ere, hiztegi ezberdinetako informazio hori zentralizatzea edo mapatzea beti ez da posible izango. Arazo hori are nabarmenagoa da baliabide gutxiko hizkuntzak tratatzen direnean. Hori dela eta, autore askok gurutzaketa-prozesuaren emaitzak hobetzeko hiztegietatik kanpo dagoen informazio berria prozesuan txertatzea proposatzen dute. Mota ugaritakoa izan daitekeen informazio gehigarri honek, hiztegiak soilik erabiliz lortzen diren emaitzen kalitatea hobetzen laguntzen du. Jarraian kanpo baliabide horiek erabiltzen dituzten teknika eta prozesuak sakonago azalduko dira.

2.2.2.1 Corpus konparagarriak *(Distributional Similarity)*

Distributional Similarity edo Antzekotasun Distribuzionalaren kalkulua arrakastaz erabiltzen da corpus konparagarrietatik terminologia elebiduna erauzteko atazatan. Prozesu honen oinarrian dagoen ideia bi hizkuntza ezberdinetako corpusetan antzeko testuinguruan agertzen diren hitzak itzulpen hautagaitzat hartzea da, beraien testuinguruek duten antzekotasun hori bi hitzen arteko distantzia semantikoarekiko proportzionala dela kontsideratuz. Beste era batera esanda, hizkuntzarteko distantzia semantikoaren eta itzulpen probabilitatearen arteko baliokidetasun bat zehazten da. Horrela, teknika hori erabiliz gurutzaketa arruntean polisemiaren ondorioz sorturiko itzulpen bikote okerrak kimatu daitezke, lantzen diren hizkuntzatako corpus konparagarrietatik ateratako informazioa erabiliz (Kaji eta Morimoto (2002); Sammer eta Soderland (2007); Tsuchiya et al. (2007); Kaji et al. (2008); Prochasson et al. (2009); Gamallo eta Pichel (2010); Shezaf eta Rappoport (2010)).

Itzulpen hautagaien antzekotasun distribuzionala kalkulatzeko teknika honen hurbilpen tradizionala erabiltzen da (Fung (1995); Rapp (1999)). *"Bag-of-word"* paradigma jarraituz, h hitz baten testuinguruak pisu zehatz bat duten hitz multzoen bidez (t_1, t_2, ..., t_n) adierazten dira. Hitz multzo hori h hitzaren inguruan finkatzen den leiho batek mugatzen du, edo beste era batera esanda, h hitzaren inguruan agertzen den hitz multzoak. Adibidez,

finkatzen den leihoa 5 elementukoa bada, h hitzaren aurreko 5 hitzak eta ondoko 5 hitzak sartzen dira testuinguruaren errepresentazioa osatzen duten hitz multzoan (± 5). Multzo hori osatzen duten hitzen pisua neurtzeko neurri ezberdinak erabiltzen dira (maiztasun absolutua, *LLR*, *Dice koefizientea*, ...). Azkenik, h hitzaren testuinguruaren errepresentazioa osatzen duten hitz guztiei pisu bat esleitu ondoren, hitz horien bidez h-ren testuingurua irudikatzen duen bektorea osatzen da.

Aztertu nahi diren hizkuntzetako hitz guztien testuinguru bektoreak sortu ondoren, hizkuntza jakin bateko hitz bakoitzeko, bere testuinguru bektorearen eta beste hizkuntzako hitz guztien bektoreen arteko antzekotasun mailak neurtzen dira bektoreen arteko kosinua kalkulatuz. Bektore horiek konparatu ahal izateko, espazio edo hizkuntza berean kokatu behar dira. Hori lortu ahal izateko, iturburu hizkuntzako bektore guztiak helburu hizkuntzara itzuli beharra dago. Itzulpen-prozesu hori aurrera eramateko lantzen diren hizkuntzak barneratzen dituzten hiztegi elebidunak erabiltzen dira. Itzulpen-prozesu hori traba handi bat izan daiteke izan ere berau egiteko behar den oinarrizko hiztegi baliabidea da pibotaje-prozesuaren emaitza izango dena, eta ondorioz kasu askotan baliabide hori ez da eskuragarri egongo. Arazo honen aurrean hainbat irtenbide planteatzen dira, besteak beste gurutzaketa egin ondoren lortzen den hiztegia zaratatsua erabiltzea (itzulpen okerrak garbitu gabe dituena) edota ziurrak diren itzulpenak soilik barneratzen dituen hiztegi bat erabiltzea (polisemiak eragiten ez dien itzulpen bikoteak).

2.2.2.2 Corpus paraleloak

Corpus paraleloak hizkuntza naturalen prozesamenduaren arloko hainbat azpiarlotan behin eta berriro erabiltzen diren baliabideak dira. Horrela, itzulpen automatikoko sistema askoren oinarrizko baliabide bihurtu dira. Corpus paraleloetako informazioaren bidez, bi hizkuntzatako hitzak estatistikoki lerrokatu daitezke, horrela hitzen arteko erlazio sendoak identifikatuz. Lerrokatze horiek hiztegi elebidunak automatikoki sortzeko prozesuan izan dezaketen erabilgarritasuna nabaria da, lerrokatzen diren hitz bikoteak itzulpen bikoteak izateko probabilitatea handia baita. Lerrokatze horiek erabilera bikoitza izan dezakete hiztegi elebidunen sorkuntza-prozesuan. Alde batetik, pibotaje tekniken bidez sorturiko hiztegietan agertzen den zarata (itzulpen okerrak) murrizteko erabili daitezke. Lehenago aipatu bezala, corpus konparagarriak ere helburu berarekin erabiltzen dira, baina corpus paraleloen bidez lorturiko emaitzen kalitatea corpus konparagarriekin lor daitekeena baino hobea da. Bestalde, lerrokatze horiek izatez itzulpen bikoteak kontsidera daitezkeenez, bikote horiek erabiliz posiblea litzateke hiztegi elebidunak hutsetik sortzea. Horrela, metodo hori pibotaje tekniken bidez sorturiko hiztegien estaldura hobetzeko ere erabili daiteke, hiztegian agertzen ez diren itzulpen bikote berriak gehitzeko aukera ematen baitu (Tsunakawa et al. (2008)).

Dena den, jakina da corpus paraleloak eskuratu edo sortzea lan nekeza dela. Hori dela eta, hizkuntza bikote askoren kasuan ezinezkoa da corpus paralelo erabilgarriak eskuratzea. Nabaria da baliabide gutxien duten hizkuntzen kasuan arazoa are larriagoa dela. Ondorioz, baliabide gutxien duten hizkuntzen kasuan corpus paraleloen erabilera oraindik ere ez da

bideragarria.

2.2.2.3 *WordNet*

WordNet[2] (Miller et al. (1990)) datu-base lexikal handi bat bezala sortu zen (hasiera batean ingelesez). Bertan, hitzak sinonimo multzotan taldekatzen dira *(Synsets)*, definizio labur eta zehatzak emanez. Multzo horien artean gainera hainbat erlazio semantiko definitzen dira, hitzen esanahian oinarrituriko sare erraldoi bat osatzen delarik. Baliabidearen agerpenak hizkuntza naturalen prozesamenduaren arloan eragin handia izan zuen, informazio linguistiko zabala eskaintzeaz gain hizkuntza guztietara hedatzeko aukera ematen duten datu egiturak erabiltzen baititu. Horrela, hizkuntza gehiagoren gehikuntza-prozesutik *EuroWordNet*[3] (Vossen (1998); Peters et al. (1998)) baliabidea sortu zen, bertan Europako hainbat hizkuntzen *WordNet* egiturak barneratzen zirelarik (Alemanera, Italiera, Gaztelera, Frantsesa, Txekiera eta Estoniera). Azken honen urratsak jarraituz, *Multilingual Central Repository*[4] baliabidea sortu zen (Agirre et al. (2007)), Euskara, Gaztelera, Katalana eta Italiera konbinatuz. Azken bi baliabide horiei esker hizkuntza ezberdinetako hitzen artean lotura semantiko sendoak definitzea lortzen da.

Aipatu berri diren baliabideak oso erabilgarriak izan daitezke hiztegi elebidunen sorkuntza automatikoan ere, izan ere, hizkuntzarteko lotura semantikoen informazioak prozesu hori izugarri sendotu dezake. Dena den, eta beste hainbat baliabiderekin gertatzen den bezala, datu-base lexikal horiek ez daude hizkuntza guztientzat eskuragarri, eta eskuragarri egonda ere, lantzen diren hizkuntza guztiak ez dira maila berean landu, oraindik ere hizkuntza batzuen presentzia beste batzuen aldean eskasa delarik. Ondorioz, baliabide gutxien duten hizkuntzen kasuan baliabide honen erabilera oraindik ere mugatua da.

Baliabide gehien duten hizkuntzak mota honetako baliabideetan duten presentzia handiagoa baliatuz, eta pibotaje-prozesuetan zubi bezala jokatzen duten hizkuntzak normalean hizkuntza handiak izaten direla jakinik (loturarik ez duten hizkuntzak erlazionatzeko zubi hizkuntza egokienak baliabide gehien dutenak izaten baitira), István eta Shoichi (2009) lanean zubi bezala erabiltzen den hizkuntzaren kasuan (ingelesa) *WordNet* baliabideak eskaintzen duen informazio gehigarria erabiltzea proposatzen da, horrela pibotaje-prozesuetan erabiltzen den informazio lexikoaz gain informazio semantiko gehigarri bat ere erabiltzeko. Informazio honen erabilerak pibotaje teknika tradizionalek dituzten hainbat arazo gainditzen lagunduko luke, izan ere informazio semantikoak bikote okerrak baztertzeaz gain bikote berriak sortzeko aukera ematen du (datu-base lexikal horietan erabiltzen diren lotura semantiko ezberdinak aplikatuz; sinonimia esaterako).

2.2.2.4 Bestelako baliabideak

[2] http://wordnet.princeton.edu/
[3] http://www.illc.uva.nl/EuroWordNet/
[4] http://adimen.si.ehu.es/web/MCR/

Sarearen hazkunde etengabearen ondorioz, gaur egun sarean hainbat eta hainbat baliabide linguistiko erabilgarri aurki daitezke (*Wikipedia*, *Wiktionary*, ...). Baliabide horiek edonorentzat eskuragarri daude, eta erabiltzaile bakoitzak bere beharretarako erabili ditzake. Horrela, baliabide horietariko batzuk oso erabilgarriak izan daitezke pibotaje tekniken bidez sorturiko hiztegi elebidunetako itzulpen bikote okerrak kimatzeko, eta are gehiago, gurutzaketa bidez lortu ezin izan diren itzulpen bikote berriak sortzeko, nahiz eta azken hori lan honetako helburuetatik kanpo geratzen den. Izan ere, baliabide horiek hizkuntza ugarirekin lan egiteko aukera eskaintzen dute. Zoritxarrez, hainbat hizkuntza handiren kasuan informazio erabilgarri izugarri egon arren, baliabide gutxien duten hizkuntzen kasuan mota honetako baliabideetatik atera daitekeen informazioa ez da hain erabilgarria, kopuruz askoz gutxiago dagoelako eskuragarri eta kalitatez ere hizkuntza handienen mailara heltzen ez delako. Beraz, lan honetan planteatzen den baliabide gutxiko hizkuntzen testuinguruan oraindik ere ezin dira baliabide erabilgarriak kontsideratu, nahiz eta aipatu bezala egoera hobetzen ari den baliabide gutxien duten hizkuntzen kasuan ere. Dena den, komenigarria da baliabide horietatik garrantzitsuenak aipatzea.

2.2.2.4.1 *Wikipedia*

Wikipedia[5] entziklopedia eleanitza, web-oinarriduna eta eduki askekoa da, bertako artikuluak Internetera konektatutako edozeinek idatz edo editatu ditzakeelarik. Ekimen hori 2.001 urtean abiatu zen eta gaur egun 200 hizkuntza baino gehiagotako edukiak biltzen ditu. Eduki eleanitz guztiak elkarren artean lotuak daude lotura ezberdinen bidez, horrela hizkuntzarteko sare erraldoi bat bezala erabil daitekeelarik. Hala ere, hizkuntza guztiak ez dira maila berean landu, mundu osoko boluntarioek sortutako sarrerak biltzen dituenez hiztun edota erabiltzaile gehien duten hizkuntzak askoz ere presentzia handiagoa baitute (adibide batzuk ematearren, ingelesezko artikuluak 4.000.000 inguru dira, euskarazkoak 100.000 inguru eta galegozkoak berriz 10.000 inguru). Dena den, nahiz eta hizkuntza batzuen presentzia oraindik ere ez hain nabaria izan, baliabide honen erabilgarritasuna izugarria da hainbat eta hainbat alorretan. Hiztegi elebidunen sorkuntza automatikoaren arloan ere, hainbat autorek *Wikipedia*-ko edukiak nola ustiatu daitezkeen aztertu dute (Erdmann et al. (2008b,a); Yu eta Tsujii (2009); Rohit eta Tandon (2010)). Horrela, hainbat autorek (Erdmann et al. (2008b,a); Yu eta Tsujii (2009)) bertako artikuluen artean finkatzen diren lotura mota ezberdinak erabiltzea planteatzen dute itzulpen bikote posibleak identifikatzeko:

- **Hizkuntzarteko loturak:** Artikulu beraren hizkuntza ezberdinetako bertsioak lotzeko erabiltzen diren link-ak dira. Mota honetako loturen bidez lotzen diren artikuluak elkarren itzulpenak izan ohi dira.

- **Berbideratze orriak:** Eduki gabeko orriak dira, eta beste artikulu batzuetarako loturak bakarrik eskaintzen dituzten. Orri horien helburua antzekotasun handia (ka-

[5] http://www.wikipedia.org/

su ohikoena sinonimia da) duten artikuluen artean loturak ezartzea da (Adibidez, ingelesezko *"phone"* eta *"telephone"* artikuluen artean).

- **Testu loturak:** Artikulu zehatz bateko edukietatik *Wikipedia*-ko beste artikulu batzuetara egiten diren erreferentziak dira. Zenbait kasutan lotura link-a adierazteko erabilitako testua eta helburu artikuluaren arteko konparaketa erabilgarria izan daiteke besteak beste genero edota numero bereizketak egiteko (adibidez, zientzia artikulu batean *"protoiak"* hitzak *"protoi"* artikulua erreferentziatu dezake, *"protoiak"* eta *"protoi"* hitzen arteko nolabaiteko lotura bat existitzen dela jakinaraziz).

Horrela, egitura horiek erabiliz itzulpen bikote berriak identifikatzeko hizkuntzarteko loturak erabiltzea proposatzen dute. Bestalde, emaitzen estaldura handitzeko, berbideratze orrietako eta testu loturetako informazioa gehigarri bezala erabiltzea proposatzen dute, bi egitura horien bidez sinonimia, hiperonimia edota hiponimia bezalako erlazioak definitu baitaitezke sarrera ezberdinen artean horrela emaitzen estaldura hobetzeko.

Bestalde, beste autore batzuk (Rohit eta Tandon (2010)) *Wikipedia*-n aurki daitezkeen egitura errepikakorrak (*Infobox* delakoak edota artikuluaren laburpen bezala erabiltzen den testuko lehen paragrafoa adibidez) erabiltzea proposatzen dute itzulpen bikoteak erauzteko, egitura horiek hizkuntza ezberdinetan errepikakorrak baitira.

2.2.2.4.2 *Wiktionary*

Wictionary[6] entziklopedia eleanitza, web-oinarriduna eta eduki askekoa da, bertako sarrerak Internetera konektatutako edozeinek idatz edo editatu ditzakeelarik. *Wikipedia*-ren "kide" lexikoa izateko sortu zen 2.002 urtean, eta gaur egun, thesaurus-ak, errima gidak, esamolde zerrendak eta beste hainbat baliabide eskaintzen ditu hainbat hizkuntzatarako. Guztira 450 hizkuntza baino gehiagoko sarreren informazioa du bere baitan, nahiz eta *Wikipedia*-rekin gertatzen den bezala hizkuntza guztiak ez dauden maila berean integratuak (adibide batzuk ematearren, txinerazko sarrerak 1.000.000 inguru dira, euskarazkoak 10.000 inguru eta danierazkoak berriz 1.000 inguru). Hizkuntza guztiek integrazio maila bera izango balute nahikoa litzateke bertako datuak zuzenean esportatzea hiztegi elebidun berriak lortzeko. Hortaz, baliabide gutxien duten hizkuntzen presentzia handitzen joan ahala *Wiktionary* baliabidea hiztegi elebidunak automatikoki sortzeko prozesuan erreferentziazko baliabide bihur daiteke.

[6] http://www.wiktionary.org/

3 Baliabideak

Aurreko atalean hiztegi elebidunen pibotaje bidezko sorkuntzan gaur egun arte landu diren hurbilpen eta proposamen guztiak aurkeztu dira. Lan honetan zehar, aurrekarietan azalduriko hainbat teknika aplikatuko dira euskarara barneratzen duten hiztegi elebidunak sortzeko eta euskararen kasuan teknika horiek duten erabilgarritasuna estimatzen saiatzeko. Helburu hori lortu ahal izateko, mota ezberdinetako baliabideak erabiliko dira. Zehazki, lan honetan erabiliko diren baliabideak hiru multzotan banatu daitezke: hiztegi gurutzaketa burutzeko softwarea *(Apertium)*, hiztegi baliabideak eta corpus baliabideak. Erabiliko diren hizkuntzak euskara, gaztelera, txinera eta ingelesa (pibote edo zubi hizkuntza bezala) izanik, beharrezkoa izango da lau hizkuntza horiekin esperimentuak burutu ahal izateko baliabide minimoak eskuratzea. Hurrengo puntuetan baliabide horien deskribapena eta propietateak azalduko dira.

3.1 *APERTIUM*

Lan honetan, hiztegi elebidunen arteko gurutzaketak automatikoki egiteko *Apertium*[7]-ek eskaintzen dituen hainbat baliabide erabili dira. *Apertium* Alacanteko unibertsitateko hizkuntza eta sistema informatikoen saileko *TRANSDUCENS*[8] ikerketa taldeak sortutako kode irekiko itzulpen automatikorako plataforma bat da eta hasiera batean gertuko hizkuntzak landu ahal izateko sortu bazen ere, aurrerapen ugari egin dira hizkuntza dibergenteagoak ere landu ahal izateko (gaztelera eta ingelesa esaterako). Helburu hori betetzeko, plataforma honek hainbat baliabide eskaintzen ditu:

- Hizkuntzekiko independentea den itzulpen tresna bat.
- Bi hizkuntzen arteko itzulpenak sortu ahal izateko oinarrizkoak diren datu linguistikoak kudeatzeko beharrezko tresna multzoa.
- Hainbat eta hainbat hizkuntza bikoterako datu linguistikoak.

Horrela, oinarrizko datu linguistikoak kudeatzea helburu duten tresnen artean, *apertium-dixtools*[9] baliabide multzoa eskaintzen da. Honek, besteak beste honako funtzionalitateak eskaintzen ditu:

- **cross**: Bi hiztegi elebidun gurutzatu XML iturburu fitxategiak erabiliz.
- **cross-param *(Crossdics)***: Bi hiztegi elebidun gurutzatu komando lerroak erabiliz.
- **merge-morph *(Merge dictionaries)***: Bi hiztegi morfologiko (elebakar) bateratu.
- **equiv-paradigms**: Paradigma baliokideak aurkitu eta erreferentziak eguneratu.

[7] http://www.apertium.org/
[8] http://transducens.dlsi.ua.es/
[9] http://wiki.apertium.org/wiki/Apertium-dixtools

- **list *(Dictionary reader)***: Hiztegi sarrerak zerrendatu.
- **dix2trie**: Sortu *Trie*[10] *(Prefix Tree)* formatuko hiztegi elebidun bat existitzen den hiztegi elebidun batetik abiatuz.
- **dix2tiny**: Sortu plataforma mugikorretarako *(j2me, palm)* datuak hiztegi elebidunetatik abiatuz.
- **reverse-bil**: Hiztegi elebiduna itzulbiratu.
- **sort *(Sort a dictionary)***: Hiztegiak ordenatu eta kategoria kontuan hartuz taldekatu.
- **format**: Hiztegiei formatua eman, aukera generiko batzuetan oinarrituz.
- **fix**: Hiztegiak konpondu (errepikatuak ezabatu, zuriuneak transformatu, ...).

Funtzio horiek oso erabilgarriak dira hiztegi elebidunak kudeatu eta landu ahal izateko. Horien artean, hiztegi elebidunak gurutzatuz hiztegi elebidun berriak sortzeko balio duten *cross* eta *cross-param* funtzioak daude. Bien funtzionamendua berdina da oinarrian, aldatzen den gauza bakarra abiapuntu bezala erabiltzen diren fitxategiak direlarik. Jarraian funtzio horien erabilera sakonago azalduko da.

3.1.1 Gurutzaketa funtzioak (*cross* eta *cross-param*)

Cross eta *cross-param apertium-dixtools* baliabide multzoaren baitan aurki daitezkeen bi funtzio dira, horien helburua hiztegi elebidunak gurutzatuz hizkuntza konbinaketa berrietarako hiztegi elebidun berriak sortzea delarik. Horrela, *A*, *B* eta *C* hiru hizkuntza ezberdin izanik, *A*-*B* eta *B*-*C* hiztegi elebidunetatik abiatuz *A*-*C* hiztegi elebiduna automatokiko sortzeko gai dira. Prozesu hori burutu ahal izateko, bi funtzio horiek gutxienez XML formatuan kodetzen diren honako baliabideak jaso behar dituzte erabiltzailearengandik:

- 2 Hiztegi elebidun
- 2 Hiztegi elebakar
- *Cross Model* dokumentua

Jarraian horietako baliabide bakoitza deskribatzen da.

[10] http://en.wikipedia.org/wiki/Trie

3.1.1.1 Hiztegi elebidunak

Gurutzaketa funtzioek gurutzaketa-prozesuan erabiltzen dituzten hiztegi elebidunak XML-n kodetzen dira. Horretarako, hiztegi elebidun bakoitza honako etiketen bidez adierazten da:

- **dictionary**: Hiztegi osoaren oinarri marka.
- **sdefs**: Erabiliko den lexikoaren kategoria gramatikal guztiak definitzen dira bertan. Kategoria bakoitzeko *sdef* etiketa erabiliko da.
- **section**: Hiztegi elebiduneko sarrerak definitzen dira bertan. Sarrera bakoitzeko *e* etiketa bat definitzen da.

Bestalde, fitxategiaren edukiaz gain fitxategiaren izenak ere egitura finko bat erakutsi behar dute beti. *A*, *B* eta *C* hizkuntzak jatorrizkoa, pibote edo zubi hizkuntza eta helburu hizkuntza izanik hurrenez hurren, hiztegi elebidunen fitxategi izenek ondoko formatua mantendu beharko dute beti:

- **A-B hiztegia**: *apertium-B-A.B-A.dix*
- **B-C hiztegia**: *apertium-B-C.B-C.dix*

Hiztegi elebidun horien izenenean adierazten den bezala, bi hiztegiek komunean duten hizkuntzak ezkerreko aldean agertu behar du beti, hau da, gurutzaketan parte hartu behar duten hiztegi elebidunak $B \rightarrow A$ eta $B \rightarrow C$ izango dira. Horretarako, *apertium-dixtools* paketeak barneratzen duen *reverse-bil* funtzioa aplikatu daiteke itzulbiratu nahi den hiztegiaren gainean, edo bestela, gurutzaketa funtzioa exekutatzen den unean adierazi parametro bezala pasatako hiztegi elebidunaren alderantzizko bertsio bat erabili behar dela. 3.1 irudian hiztegi elebidun baten adibide bat ikus daiteke.

```
<dictionary>
  <sdefs>
    <sdef n="ize"/>
  </sdefs>
  <section id="main" type="standard">
  ...
  <e><p><l>ebakitzaile<s n="ize"/></l><r>cutting<s n="ize"/></r></p></e>
  <e><p><l>ebakitzaile<s n="ize"/></l><r>secant<s n="ize"/></r></p></e>
  <e><p><l>ebakortz<s n="ize"/></l><r>incisor<s n="ize"/></r></p></e>
  ...
  </section>
</dictionary>
```

Irudia 3.1: Hiztegi elebidun baten adibidea

```
<dictionary>
   <sdefs>
     <sdef n="ize"/>
     <sdef n="sg"/>
     <sdef n="pl"/>
   </sdefs>
   <pardef n="ak_plurala">
     <e><p><l/><r><s n="ize"/><s n="sg"/></r></p></e>
     <e><p><l>ak</l><r><s n="ize"/><s n="pl"/></r></p></e>
   </pardef>
   <section>
     ...
     <e><i>ebakitzaile</i><par n="ak_plurala"/></e>
     <e><i>ebakortz</i><par n="ak_plurala"/></e>
     ...
   </section>
</dictionary>
```

Irudia 3.2: Hiztegi elebakar baten adibidea

3.1.1.2 Hiztegi elebakarrak

Hiztegi elebidunez gain, gurutzaketa burutu ahal izateko gurutzaketa funtzioek hiztegi elebidun horietan agertzen diren jatorri eta helburu hizkuntzetako lexiko guztia definitzen duten bi hiztegi elebakar morfologiko ere behar dituzte. Hiztegi horiek ere XML-n kodetzen dira, ondokoak direlarik horiek definitzeko erabili beharreko etiketak:

- **dictionary:** Hiztegi osoaren oinarri marka.
- **sdefs:** Erabiliko den lexikoaren kategoria gramatikal guztiak definitzen dira bertan. Kategoria bakoitzeko *sdef* etiketa erabiliko da.
- **pardefs:** Paradigmak definitzen dira bertan. Paradigmek hitzen portaera nolakoa den definitzeko balio dute. Paradigma horiek besteak beste sarrera zehatz batzuek generoaren edo numeroaren arabera zein portaera duten adierazteko balio dute. Adibidez, 3.2 irudian ageri den adibidean definitzen den paradigmak numeroa plurala den kasuetan sarrerari *"-ak"* amaiera gehitu behar zaiola definitzen da, sarrerak paradigma hori betetzen duela definitzen den kasuetan. Paradigma bakoitza *pardef* etiketaren bidez adierazten da.
- **section:** Hiztegi elebakarreko sarrerak definitzen dira bertan. Sarrera bakoitzeko *e* etiketa bat definitzen da.

Bestalde, fitxategiaren edukiaz gain fitxategiaren izenak ere egitura finko bat erakutsi behar dute beti. *A*, *B* eta *C* hizkuntzak jatorrizkoa, pibote edo zubi hizkuntza eta helburu hizkuntza izanik hurrenez hurren, hiztegi elebakarren fitxategi izenek ondoko formatua mantendu beharko dute beti:

HAP masterra

- **A-B hiztegia:** *apertium-B-A.A.dix*
- **B-C hiztegia:** *apertium-B-C.C.dix*

Hiztegi elebidunen izenetan gertatzen den moduan, hiztegi elebakarretan ere zubi edo pibote moduan jokatzen duen hizkuntza kokatzen da ezkerraldean. Puntuaren ondoren hiztegi elebakarraren hizkuntza zehazten da. 3.2 irudian hiztegi elebakar baten adibide bat ikus daiteke.

3.1.1.3 *Cross Model* fitxategia

Hizkuntza ezberdinen arteko gurutzaketak egin ahal izateko, gurutzaketa funtzioei gurutzaketa ekintzak *(cross actions)* direlakoak zehazteko aukera ematen da. Gurutzaketa ekintza horiek patroi batzuek eta patroi horiek betetzen diren kasuetan jarraitu beharreko ekintzek osatzen dituzte. Fitxategi hori definitzeko ondoko etiketa multzoa erabiltzen da:

- **cross-model:** Gurutzaketa ekintza fitxategiaren oinarri etiketa.
- **cross-action:** Parekatze ekintza bat definitzeko erabiltzen da.
- **description:** Parekatze ekintza baten deskribapena emateko erabiltzen da.
- **pattern:** Parekatze ekintza zehatz bat exekuta dadin bilatzen den parekatze patroia definitzen da.
- **action-set:** Parekatze patroi bat betetzen denean bete beharreko ekintza multzoa definitzeko erabiltzen da.
- **action:** Ekintza multzoan *(action-set)* barneratzen den parekatze ekintza bakoitza definitzeko erabiltzen da.

3.3 irudian fitxategi mota honen oinarrizko adibide bat ikus daiteke.

3.1.2 Gurutzaketa-prozesua

Gurutzaketa burutzeko beharrezkoak diren baliabide guztiak eskura izanda, gurutzaketa-prozesua bi eratara burutzeko aukera ematen da: *LRD* fitxategia *(Linguistic Resources Document)* erabiliz (*cross* funtzioa), edota hori erabili gabe (*cross-param* funtzioa). *LRD* fitxategia prozesuan parte hartzen duten parametro nahiz baliabide guztiak definitzeko erabiltzen da, eta beraz, hori erabiltzen ez den kasuetan bertan adierazten den informazio guztia beste era batera adierazi beharra dago. Jarraian gurutzaketa-prozesua burutzeko *LRD* fitxategia erabiliz eta hori erabili gabe eman beharreko urratsak definitzen dira.

HAP masterra

```
<cross-model>
  <cross-action id="default" a="ebenimeli">
    <description>Default pattern</description>
    <pattern>
      <e>
        <p>
          <l>$lemmaA<v n="cat"/><t n="tailA"/></l>
          <r>$lemmaB<v n="cat"/><t/></r>
        </p>
      </e>
      <e>
        <p>
          <l>$lemmaB<v n="cat"/><t/></l>
          <r>$lemmaC<v n="cat"/><t n="tailC"/></r>
        </p>
      </e>
    </pattern>
    <action-set>
      <action>
        <e>
          <p>
            <l>$lemmaA<v n="cat"/><t n="tailA"/></l>
            <r>$lemmaC<v n="cat"/><t n="tailC"/></r>
          </p>
        </e>
      </action>
    </action-set>
  </cross-action>
</cross-model>
```

Irudia 3.3: Oinarrizko *Cross Model* fitxategi baten adibidea

3.1.2.1 *LRD* fitxategia erabiliz *(cross)*

Funtsean, *LRD* fitxategia baliabide linguistiko multzo bat definitzeko erabiltzen den XML formatudun fitxategi bat da. Baliabide horiek besteak beste hiztegiak, corpusak edota beste *LRD* fitxategi batzuetarako erreferentziak izan daitezke. Fitxategia bloke ezberdinetan banatzen da, bakoitzean informazio mota ezberdina zehazten delarik. Bloke horietan definitzen diren hiztegi baliabideak erabiliz, *cross* funtzioa hiztegi elebidun berriak sortzeko gai da. 3.5 irudian eta 3.6 irudian *LRD* fitxategi baten egitura eta erabilera konkretu baten adibidea ikus daitezke hurrenez hurren. Jarraian *LRD* fitxategi bat osatzen duten blokeak sakonago azalduko dira.

3.1.2.1.1 *RESOURCE* blokea (Baliabide blokea)

LRD fitxategian definitzen den baliabide bakoitza propietate multzo baten bidez definitzen da (Ikus 3.4 irudia).

HAP masterra

```
<resource>
   <property name="name" value="apertium-es"/>
   <property name="type" value="mon"/>
   <property name="sl" value="es"/>
   <property name="for-crossing" value="yes"/>
   <property name="src" value="apertium-es-ca.es.dix"/>
   <property name="version" value="stable"/>
</resource>
```

Irudia 3.4: *LRD* fitxategi baten baliabide blokearen adibidea

Propietate horien balio posibleak honakoak izan daitezke:

- **name:** Baliabidearen izena zehazteko erabiltzen da.
- **type:** Baliabidearen mota definitzeko erabiltzen da. Bere balio posibleak honakoak izan daitezke:
  - *mon:* Hiztegi morfologikoa
  - *bil:* Hiztegi elebiduna
  - *crp:* Corpusa
  - *lrd:* Beste *LRD* fitxategi baterako erreferentzia
  - *cross-model:* Gurutzaketa ekintzak definitzen dituen fitxategia
- **sl:** Jatorri hizkuntza (hiztegi morfologikoetan eta hiztegi elebidunetan)
- **tl:** Helburu hizkuntza (hiztegi elebidunetan)
- **src:** Iturburu fitxategiaren URLa edo PATH
- **version:** Eskaintzen den baliabidearen uneko bertsioa

3.1.2.1.2 *RESOURCE-SET* blokea (Baliabide multzo blokea)

LRD fitxategiak baliabide blokean definitzen diren erako baliabideen multzoak definitzeko aukera ematen du (ikus 3.5 irudia). Multzokatze horiek egokiak izan daitezke adibidez hizkuntza bikote konkretu bateko baliabideak antolatzeko.

3.1.2.1.3 Exekuzioa

Gurutzaketa-prozesurako beharrezkoak diren baliabide guztiak definitzen dituen *LRD* fitxategia sortu ostean, nahikoa da ondoko komandoa exekutatzea, parametro bezala bi balio emanez: *LRD* fitxategia eta hizkuntza bikotearen kodea (*A* eta *B* gure hizkuntzak izanik, *A-B*):

$ apertium-dixtools cross -f my-linguistic-resources.xml A-B

HAP masterra

```
<ling-resources>
   <name>...</name>
   <description>...</description>
   <resource>
      <property name="..." value="..."/>
      <property name="..." value="..."/>
      ...
   </resource>
   <resource-set>
      <name>...</name>
      <description>...</description>
      <resource>
         <property name="..." value="..."/>
         ...
      </resource>
      <resource>
         <property name="..." value="..."/>
         ...
      </resource>
      ...
   </resource-set>
   <resource>
      <property name="..." value="..."/>
      <property name="..." value="..."/>
      ...
   </resource>
   ...
</ling-resources>
```

Irudia 3.5: *LRD* fitxategi baten egitura orokorra

3.1.2.2 *LRD* fitxategia erabili gabe *(cross-param)*

LRD fitxategia erabiltzen ez den kasuetan, bertan definitu ohi diren baliabide guztiak komando lerrotik pasatu beharra dago. Hortaz, *A*, *B* eta *C* hurrenez hurren jatorri hizkuntza pibote edo zubi hizkuntza eta helburu hizkuntza izanik, komandoari parametro bezala pasatu beharreko parametro zerrenda ondokoa izango litzateke (2.1.1 eta 2.1.2 puntuetan azaldu bezala, baliabide horien izenak formatu konkretu bat jarraitu behar dute beti):

- ***A-B* hiztegi elebiduna:** *apertium-B-A.B-A.dix*
- ***B-C* hiztegi elebiduna:** *apertium-B-C.B-C.dix*
- ***A* hiztegi elebakarra:** *apertium-B-A.A.dix*
- ***C* hiztegi elebakarra:** *apertium-B-C.C.dix*

HAP masterra

```
<!-- Linguistic resources-->
<ling-resources>
   <name>My linguistic resources</name>
   <description>My linguistics resources: morphological and bilingual dictionaries, cross models, corpora,
etc.</description>
   <resource-set>
      <name>My linguistic resources to get English-Spanish language pair.</name>
      <description>A description of this resource set</description>
      <!-- cross model es-ca-en -->
      <resource>
         <property name="name" value="cross-model-es-ca-en"/>
         <property name="type" value="cross-model"/>
         <property name="sl" value="es"/>
         <property name="tl" value="en"/>
         <property name="for-crossing" value="yes"/>
         <property name="src" value="cross-model-es-ca-en.xml"/>
         <property name="version" value="stable"/>
      </resource>
      <!-- 'es' morphological dictionary -->
      <resource>
         <property name="name" value="apertium-es"/>
         <property name="type" value="mon"/>
         <property name="sl" value="es"/>
         <property name="for-crossing" value="yes"/>
         <property name="src" value="apertium-es-ca.es.dix"/>
         <property name="version" value="stable"/>
      </resource>
      <!-- 'en' morphological dictionary -->
      <resource>
         <property name="name" value="apertium-en"/>
         <property name="type" value="mon"/>
         <property name="sl" value="en"/>
         <property name="for-crossing" value="yes"/>
         <property name="src" value="apertium-en-ca.en.metadix"/>
         <property name="version" value="stable"/>
      </resource>
      <!-- 'en-ca' bilingual dictionary -->
      <resource>
         <property name="name" value="apertium-en-ca"/>
         <property name="type" value="bil"/>
         <property name="sl" value="en"/>
         <property name="tl" value="ca"/>
         <property name="for-crossing" value="yes"/>
         <property name="src" value="apertium-en-ca.en-ca.dix"/>
         <property name="version" value="stable"/>
      </resource>
      <!-- 'es-ca' bilingual dictionary -->
      <resource>
         <property name="name" value="apertium-es-ca"/>
         <property name="type" value="bil"/>
         <property name="sl" value="es"/>
         <property name="tl" value="ca"/>
         <property name="for-crossing" value="yes"/>
         <property name="src" value="apertium-es-ca.es-ca.dix"/>
         <property name="version" value="stable"/>
      </resource>
   </resource-set>
   <!-- Single corpus file -->
   <resource>
      <property name="name" value="corpus-es"/>
      <property name="type" value="corpus"/>
      <property name="sl" value="es"/>
      <property name="src" value="corpus-es.crp"/>
   </resource>
</ling-resources>
```

Irudia 3.6: *LRD* fitxategi baten erabilera konkretu baten adibidea

3.1.2.2.1 Exekuzioa

Baliabide sorta horretaz gain, hiztegi elebidunen kasuan hiztegien norabidea adierazi beharra dago, hau da, hiztegia dagoen bezala gurutzatu behar den *(-n)* edo itzulbiratu beharra dagoen *(-r)*. Izan ere, gogoratu 2.2.1 puntuan azaldu bezala pibote edo zubi hizkuntzak beti ezkerreko aldean agertu beharko duela beti hiztegietan. Beraz, erabiliko den baliabide multzoa eta hiztegi elebidunen norabideak kontuan hartuz, gurutzaketa burutzea ahalbidetzen duen komandoa ondokoa da:

$ apertium-dixtools cross-param apertium-B-A.A.dix -n apertium-B-A.B-A.dix -n apertium-B-C.B-C.dix apertium-B-C.C.dix

3.1.3 Gurutzaketa-prozesuaren emaitza

Aurreko puntuan definitutako komandoaren bidez *dix* izeneko direktorio berri bat sortzen da, honen barruan gurutzaketak eman dituen emaitza fitxategiak sortzen direlarik. Emaitza hori osotara 12 fitxategik osatzen dute. Fitxategi guztiak XML formatuan daude, eta bakoitzak informazio mota ezberdin bat gordetzen du. Jarraian sortzen diren fitxategi guztien deskribapen labur bat agertzen da, prozesuan parte hartu duten hizkuntzak A, B eta C (jatorri hizkuntza, pibote edo zubi hizkuntza eta helburu hizkuntza hurrenez hurren) direla suposatuz:

- **apertium-A-C.C-crossed.dix:** Gurutzaketa egin ondoren gurutzatuak izan diren C-ko sarrera hitzen hiztegi elebakarra (ordenatua).
- **apertium-A-C.A-crossed.dix:** Gurutzaketa egin ondoren gurutzatuak izan diren A-ko sarrera hitzen hiztegi elebakarra (ordenatua).
- **apertium-A-C.A-C-crossed.dix:** Gurutzaketa egin ondoren lorturiko A-C hiztegi elebiduna (gurutzaketaren emaitza nagusia).
- **consistent-bilAB.dix:** Gurutzaketan erabili diren hitzen A-B hiztegi elebiduna.
- **consistent-bilBC.dix:** Gurutzaketan erabili diren hitzen B-C hiztegi elebiduna.
- **consistent-monA.dix:** Gurutzaketan erabili diren hitzen A hizkuntzako hiztegi elebakarra.
- **consistent-monC.dix:** Gurutzaketan erabili diren hitzen C hizkuntzako hiztegi elebakarra.
- **not-common-apertium-B-C.B-C.dix:** Gurutzatu gabe geratu diren hitzen B-C hiztegi elebiduna.
- **not-common-apertium-B-C.C.dix:** Gurutzatu gabe geratu diren hitzen C hiztegi elebakarra.

HAP masterra

- **not-common-apertium-B-A.B-A.dix:** Gurutzatu gabe geratu diren hitzen *A-B* hiztegi elebiduna.
- **not-common-apertium-B-A.A.dix:** Gurutzatu gabe geratu diren hitzen *A* hiztegi elebakarra
- **patterns-not-detected.xml:** Gurutzaketa-prozesuan detektatu ez diren patroiak agertzen dira bertan. Fitxategi honetako informazioa oso lagungarria izan daiteke *Cross Model* fitxategian definitzen diren patroiak hobetzeko (gurutzaketa ekintza zehatzagoak definituz).

Beraz, prozesu guztiaren emaitza den *A-C* hiztegi elebiduna *apertium-A-C.A-C-crossed.dix* fitxategian gordetzen da, 2.1.1 puntuan definitutako XML formatuan.

3.2 Hiztegi baliabideak

Aurreko atalean azaldu bezala, pibotaje tekniken oinarrizko ideia existitzen diren hiztegi elebidunak gurutzatuz hizkuntza bikote berriak barneratzen dituzten hiztegi elebidun berriak sortzea da. Beraz, nabarmena da prozesuaren abiapuntua hiztegi elebidunak direla. Lan honetan erabili diren hiztegi elebidunek ez dute sarreren definiziorik ematen, hitz bakarreko itzulpenak baizik, aurreko atalean azaldu bezala pibotaje teknikak hitz bakarreko itzulpenak dituzten hiztegietan erabiltzen baitira. Pibote edo zubi hizkuntza modura ingelesa erabili da esperimentu guztietan, eta ondorioz, emaitzen ebaluaziorako erabili diren hiztegietan izan ezik gainerako guztietan lantzen den hizkuntza bat ingelesa da, bigarren hizkuntza delarik aldatzen dena (euskara, gaztelera edota txinera). Hiztegi elebidun horiek beraien artean konparatu ahal izateko, oinarrizko ezaugarri komun batzuk definitu daitezke:

- **Sarrera kopurua:** Hiztegi elebidunak duen sarrera kopurua neurtzen du.
- **Itzulpen kopurua:** Hiztegi elebidunak duen itzulpen kopurua neurtzen du.
- **Anbiguotasun maila:** Hiztegi elebidunak sarrera bakoitzeko bataz beste ematen duen itzulpen kopurua neurtzen du.

3.1 taulan lan honetan erabili diren hiztegi elebidun guztien oinarrizko ezaugarri horiek ikus daitezke. Jarraian, erabilitako hiztegi bakoitzaren deskribapen labur bat egingo da.

Hiztegia	Sarrera kopurua	Itzulpen kopurua	Anbiguotasun maila
EU-EN	17.672	43.021	2,43
EN-ES	16.326	38.128	2,33
EU-ES	57.334	138.579	2,42
EN-ZH	37.313	63.899	1,71

Taula 3.1: Erabilitako hiztegi elebidunen oinarrizko ezaugarriak

3.2.1 *EU-EN* hiztegia

Euskara eta ingelesa barneratzen dituen hiztegi elebidun bezala Elhuyar Euskara-Ingelesa hiztegia erabili da. Hiztegi honek euskarazko 17.672 sarrera ditu eta ingelesezko 43.021 itzulpen (sarrera bakoitzeko bataz beste 2,43 itzulpen). Hiztegiak ondoko informazioa ematen du sarrera bakoitzeko:

- **Adiera ezberdinak:** Sarrera bakoitzaren adiera ezberdinak bereizten ditu zenbaki indize bat erabiliz. Adiera bakoitzeko itzulpen bat edo gehiago ematen dira.
- **Kategoriak:** Sarrera bakoitzaren kategoria gramatikala ematen du. Guztira 41 kategoriaz osatua dagoen sailkapen sistema bat erabiltzen du.

Hiztegia testu formatu arruntean kodetua dago eta sarrera bakoitzaren itzulpen bakoitza (adiera bakoitzaren aldaera bakoitza) lerro batean ematen da ondoko formatuan:

sarrera adiera_zenbakia kategoria itzulpena

Adibidez:

abiarazle 1 iz. initiator

3.2.2 *EN-ES* hiztegia

Ingelesa eta gaztelera barneratzen dituen hiztegiak ingelesezko 16.326 sarrera ditu eta gaztelerazko 38.128 itzulpen (sarrera bakoitzeko bataz beste 2,33 itzulpen). Hiztegiak ondoko informazioa ematen du sarrera bakoitzeko:

- **Adiera ezberdinak:** Sarrera bakoitzaren adiera ezberdinak bereizten ditu. Horretarako, sarrera bakoitzaren adiera bakoitzari dagozkien itzulpen guztiak lerro berean erakusten ditu, komaz separaturik.
- **Itzulpenen izaera:** Sarrera bakoitzeko izaera ugaritako itzulpenak ematen dira. Izaera mota ezberdin hauek bereizten dira:
  - *T:* Itzulpen arrunta.
  - *C:* HAUL (hitz anitzeko unitate lexikala) motako itzulpena.
  - *I:* Idiom motako itzulpena.
- **Kategoriak:** Sarrera bakoitzaren adiera bakoitzeko kategoria gramatikala ematen du. Guztira 15 kategoriaz osatua dagoen sailkapen sistema bat erabiltzen du.
- **Klase semantikoak:** Sarrera bakoitzaren adiera bakoitzeko klase semantikoa ematen du. Klase honen bidez sarrerak esanahiaren arabera sailkatzen dira. Guztira 72 klasez osatua dagoen sailkapen sistema bat erabiltzen du.

HAP masterra

Hiztegia testu formatu arruntean kodetua dago eta sarrera bakoitzaren adiera bakoitzaren itzulpen guztiak lerro batean ematen dira ondoko formatuan (informazioak ez du kasu guztietan osorik agertu beharrik):

sarrera izaera <kategoria> <klase_semantikoa> <sarreraren_HAUL_erabilera> <sarreraren_idiom_erabilera> <itzulpena(k)>

Adibidez:

amusement T <n> < > < > < > <diversión,entretenimiento>

3.2.3 *EU-ES* hiztegia

Euskara eta gaztelera barneratzen dituen hiztegia Elhuyarreko Euskara-Gaztelera hiztegi txikia eta Elhuyarreko Euskara-Gaztelera hiztegi handia konbinatuz lorturiko hiztegia da. Guztira, euskarazko 57.334 sarrera eta gaztelerazko 138.579 itzulpen ditu (sarrera bakoitzeko bataz beste 2,42 itzulpen). Automatikoki sortutako hiztegien kalitatea ebaluatzeko erreferentzia bezala erabiltzeko sortu denez, helburu horretarako beharrezkoa den informazio minimoa bakarrik aurki daiteke bere baitan, hau da, sarrerak eta horien itzulpen posibleak. Kategoria, adiera edo klase semantikoen informaziorik ez du ematen. Hiztegia testu formatu arruntean kodetua dago eta sarrera bakoitzaren itzulpen bakoitza lerro batean ematen du:

sarrera itzulpena

Adibidez:

txakur perro

3.2.4 *EN-ZH* hiztegia

Ingelesa eta txinera barneratzen dituen hiztegi elebidun bezala *MDBG*[11] hiztegia erabili da. Hiztegi honek ingelesezko 17.699 sarrera eta txinerazko 42.994 itzulpen ditu (sarrera bakoitzeko bataz beste 2,43 itzulpen). Hiztegiak ondoko informazioa eskaintzen du sarrera bakoitzeko:

- Sarrera tradiziozko txinerazko alfabetoan adierazia.
- Sarrera txinera sinplifikatuko alfabetoan adierazia.
- Txinerazko sarreren *pinyin*[12] (txinerazko karaktereak kodeketa latinora pasatzeko sistema) aldaera.

[11] http://www.mdbg.net/chindict/chindict.php
[12] http://en.wikipedia.org/wiki/Pinyin

Hiztegia testu formatu arruntean kodetua dago eta sarrera bakoitzaren informazio guztia lerro bakarrean ematen ondoko formatuan:

sarrera (tradiziozko_txineran) sarrera (txinera_sinplifikatuan) [pinyin_aldaera] itzulpenak ("/"-z bereiziak)

Adibidez:

德律 德律风 [de2 lu:4 feng1] /telephone/

3.2.5 *EN-ZH* kontsulta hiztegia

Euskara-Txinera hiztegiaren eskuzko ebaluazioa egin ahal izateko, eta txinatar hizkuntzan ezagutza nahikorik duen giza baliabide faltaren ondorioz, ebaluazio hori burutu ahal izateko sarean eskuragarri dagoen Txinera-Ingelesa kontsulta hiztegi bat erabili da oinarrizko laguntza bezala. Hiztegiaren izena *Yellow Bridge*[13] da eta besteak beste txinerazko sarreren itzulpenak, sinonimoak eta erabilpen adibideak eskaintzen ditu. Hiztegi hori ebaluazio helburuetarako bakarrik erabiliko da.

3.3 Corpus baliabideak

Aurreko atalean azaldu bezala, eta batez ere baliabide gutxien duten hizkuntzen kasuan, pibotaje tekniken bidez sortzen diren hiztegi elebidunetan itzulpen okerrak kimatzeko corpus konparagarrietan oinarritzen den *DS (Distributional Similarity)* teknika oso erabilgarria da, corpus konparagarriak lortzea corpus paraleloak lortzea baino errazagoa baita. Lan honetan zehar, *DS* teknika behin baino gehiagotan erabiliko da, aurreko ataletan aipatu bezala gure helburua baliabide gutxien duten hizkuntzen testuinguruan lan egitea baita. Jatorri eta helburu hizkuntzak euskara, gaztelera eta txinera izanik, *DS* aplikatu ahal izateko beharrezkoa izango da hizkuntza bakoitzeko gutxienez corpus elebakar bat eskuratzea. Gainera, beharrezkoa izango da hizkuntza bikote bakoitzeko erabiltzen diren corpus elebakarrak beraien artean konparagarriak izatea. Gure kasuan, corpus horiek albisteetatik abiatuz sortu dira (kazetaritza domeinua) eta beraien arteko konparagarritasuna bermatzeko ondoko baldintzak ezarri dira:

- Corpus guztiak denbora tarte bereko albisteez osatzen dira (2.006-2.010).
- Corpus guztiak mota edo kategoria (kirola, mundua, politika,...) bereko albisteen bidez osatuta daude. Hizkuntza ezberdinetan komunak ez diren motatako albisteak (albiste lokalak adibidez) ez dira corpusetan gehitzen.

Lan honetan erabili diren corpus guztiak internetetik eskuratu eta osatuak izan dira. 3.2 taulan corpus horien tamainak ikus daitezke. Jarraian corpus horietako bakoitzaren deskribapen labur bat egingo da.

[13] http://www.yellowbridge.com/

Corpusa	Dokumentu kopurua	Hitz kopurua
EU Corpusa	150K	37M
ES corpusa	307K	77M
ZH corpusa	86K	55M

Taula 3.2: Erabilitako corpusen oinarrizko ezaugarriak

3.3.1 *EU* Corpus elebakarra

Euskarazko corpus elebakarra kazetaritza domeinuko da eta *Berria*[14] egunkariak bere web orrian publikatzen dituen albisteez osatua dago (2.006-2.010 tarteko albisteak). Guztira 37 milioi hitz inguru ditu 150.000 dokumentutan (*txt* eta *sgml* formatuan) banatuak. Corpus honetako informazioa behar bezala erabili eta prozesatu ahal izateko testu guztiak etiketatua izan dira *Eustagger* (Ezeiza et al. (1998)) etiketatzaile linguistikoaren bidez.

3.3.2 *ES* Corpusa elebakarra

Gaztelerazko corpus elebakarra kazetaritza domeinukoa da eta *Diario Vasco*[15] egunkariak bere web orrian publikatzen dituen albisteez osatua dago (2.006-2.010 tarteko albisteak). Guztira 77 milioi hitz inguru ditu 307.000 dokumentutan (*txt* eta *sgml* formatuan) banatuak. Corpus honetako informazioa behar bezala erabili eta prozesatu ahal izateko testu guztiak etiketatua izan dira *Freeling*[16] etiketatzaile linguistikoaren bidez.

3.3.3 *ZH* Corpusa elebakarra

Txinerazko corpus elebakarra kazetaritza domeinukoa da eta *Beijing Daily*[17] egunkari txinatarrak bere web orrian publikatzen dituen albisteez osatua dago (2.006-2.010 tarteko albisteak). Testu guztiak txinerazko alfabeto sinplifikatuan idatziak daude, eta guztira 55 milioi tokeneko testu masa bat osatzen dute, edukia 86.000 dokumentutan banatua dagoelarik. Corpus honetako informazioa behar bezala erabili eta prozesatu ahal izateko testu guztiak tokenizatua izan dira *Stanford Chinese Segmenter*[18] (Tseng (2005)) tokenizatzailearen bidez.

3.3.4 Corpus paralelo lagungarriak

Hiztegien sorkuntza-prozesuan zuzenean parte hartzen duten corpus konparagarriez gain, esperimentuetan zehar lan honetan lantzen diren bi hizkuntza bikoteetako bakoitzeko (Euskara-Gaztelera eta Euskara-Txinera) corpus paralelo bana ere erabiltzen da, sorturiko

[14] http://www.berria.info/
[15] http://www.diariovasco.com/
[16] http://nlp.lsi.upc.edu/freeling/
[17] http://www.bjd.com.cn/
[18] http://nlp.stanford.edu/software/segmenter.shtml

esperimentuen propietate konkretu baten azterketa egin ahal izateko (hiztegien sorkuntza-prozesuan ez dute parte hartzen). Corpus horien bidez aztertzen den faktorea itzulpen probableenen estaldura da (4, eta 5. kapitulutan sakonago azalduko da). Funtsean aztertzen diren hiztegietako sarreren itzulpenak ohikoenak edo arraroagoak diren aztertzea da corpus horien helburua. Horretarako, itzulpen bikote guztien maiztasunak erauzten dira corpus paralelotik, eta sarrera bakoitzeko maiztasun handienarekin agertzen diren itzulpenak itzulpen problabeenak bezala identifikatzen dira. Horrela, aztergaiak diren hiztegiek itzulpen bikote ohikoen horiek nola tratatzen dituzten ebaluatzea lor daiteke. Jarrain bi corpus horien deskribapen txiki bat egingo da.

3.3.4.1 *EU-ES* corpus paraleloa

Euskara-Gaztelera corpus paralelo hori kazetaritza domeinukoa eta osotara 295.026 segmentuz osatua dago. Hitz kopuruari dagokionez, euskarazko 4 milioi hitz ditu eta gaztelerazko 4,7 milioi hitz. Aurreko puntuan azaldu bezala, lan honetan zehar ebaluazio helburuetarako soilik erabiltzen da.

3.3.4.2 *EU-ZH* corpus paraleloa

Euskara-Txinera corpus paralelo hori Bibliaren itzulpenean oinarrituriko corpus bat da eta osotara 31.102 segmentuz osatua dago. Tamaina eta lexiko mugatua izan arren, sarri erabiltzen diren hiztegi sarrerak lantzeko adina informazio eskain dezake.

4 Esperimentuak

Lan honen helburu nagusia hiztegi elebidunak automatikoki eraikitzea ahalbidetzen duten pibotaje teknikak aztertzea eta probatzea da. Zehazki, baliabide gutxien duten hizkuntzen testuingurua aztertzea erabaki da, azken finean horiek baitira baliabide behar handiena dutenak eta ondorioz teknika automatiko horietatik onura gehien lortuko luketenak. Testuinguru honetan, euskara hartuko da abiapuntutzat, izan ere, nahiz eta azken urteetan aurrerapen handiak egin, baliabide kopuruaren ikuspuntutik oraindik ere hizkuntza handienetatik oso urrun dago. Teknikei dagokionez, esperimentuak *IC (Inverse Consultation)* eta *DS (Distributional Similarity)* teknikekin burutzea erabaki da, lan honen bigarren atalean azaldu bezala horiek baitira baliabide gutxien duten hizkuntzen kasuan erabilgarrienak. Horrela, aurkeztuko diren esperimentuen helburua bikoitza da. Alde batetik lortzen diren hiztegi elebidun berrien kalitatea ebaluatu nahi da, eszenatoki errealetan erabilera posiblerik ba ote duten estimatzeko. Bestetik, baliabide gutxien duten hizkuntzetara egokitzen diren pibotaje teknikak aztertu nahi dira eta horien arteko konparaketa bat burutu bakoitzaren ezaugarriak identifikatzeko.

4.1 Esperimentuen diseinu-faktoreak

Beraz, euskara jatorri hizkuntza bezala finkaturik, lan honetan egingo diren esperimentuetan pibote edo zubi hizkuntza eta helburu hizkuntza aukeratzeko irizpideak finkatu behar dira. Zubi hizkuntzaren kasuan, nabarmena da ingelesa dela aukera logikoena, izan ere hori da baliabide gehien duen hizkuntza eta ondorioz loturarik ez duten hizkuntzak lotzeko hautagaia erabilgarriena. Helburu hizkuntza aukeratzerakoan aldiz, faktore gehiago hartu behar dira kontuan. Faktore horiek jarraian deskribatzen dira.

4.1.1 Baliabideak

Nabaria da pibotaje tekniken bidez hiztegi elebidun berriak sortzeko baliabide minimo batzuk beharrezkoak direla. Zehazki eta bigarren atalean azaldu bezala, *IC* teknika erabiltzeko beharrezko baldintza da bi hiztegi elebidun eskura izatea, eta *DS* teknikaren kasuan, bi hiztegi horiez gain jatorri eta helburu hizkuntzako corpus elebakar bana ere eduki beharra dago (beraien artean konparagarriak gainera). Hiztegi elebidunen kasuan gainera, hainbat propietate komunean izatea komenigarria da:

- **Itzulpen kopuru altua:** Gurutzaketa-prozesua itzulpen bikote konkretu bateko sarrerak eta itzulpenak komunean dituzten pibote hitzetan oinarritzen denez, komenigarria da hiztegi elebiduneko sarreren itzulpen kopurua ahal den altuena izatea. Izan ere, itzulpen kopuru altu honek pibote hitz gehiago egotea suposatzen du, honek itzulpen bikote berri gehiago lortzeko probabilitatea handitzen duelarik.
- **Hiztegien antzekotasuna:** Komenigarria da gurutzaketan parte hartzen duten hiztegi elebidunek antzeko izaera izatea. Izaera hori definitzen duten ezaugarri batzuk besteak beste hauek izan daitezke:

  - *Idazkera estiloa:* aditzak jokatuak edo jokatu gabe agertzen diren, izenak singularrak edo pluralak,...
  - *Idazkera maila:* Ematen diren itzulpenak arruntak edo teknikoagoak diren.
  - *Itzulpen irizpidea:* Ematen diren itzulpenak erabilienak diren, edo zuzenenak diren (nahiz eta hain erabiliak ez izan), ...

- **Tamaina:** Hiztegiek antzeko tamaina izatea (sarrera kopuruan nahiz itzulpen kopuruan) komenigarria da, bestela gurutzaketaren emaitza tamaina txikieneko hiztegiak mugatzen baitu.

4.1.2 Hizkuntzen gertutasuna

Jakina da hizkuntza bikote batzuk beste batzuek baino gertutasun maila altuagoa dutela, arrazoi ezberdinen ondorioz (kokapen geografikoa, jatorria,...). Faktore honek pibotaje tekniken bidez sorturiko hiztegi elebidunen kalitatean eragina izan dezakeela pentsa daiteke.

4.1.3 Emaitzen ebaluazio-prozesua

Esperimentuetan lorturiko hiztegi emaitzen ebaluazio fasea ere kontuan hartu behar da helburu hizkuntza aukeratzerakoan. Emaitza bezala lortzen diren hiztegien hainbat ezaugarri azter daitezke, bakoitza mota ezberdinetako ondorioak ateratzeko erabili daitekeelarik. Ezaugarri horiek hurrengo atalean zehatzago deskribatuko diren arren, jarraian zerrendatzen dira:

- **Sarreren estaldura:** Hiztegiak lantzen dituen sarrerek hizkuntzako lexikoa zenbatean estaltzen duten estimatzen du.

- **Itzulpenen estaldura:** Sarrera konkretu baten adiera posible guztietatik eta horietako bakoitzaren aldaera guztietatik eskaintzen dituenen estaldura estimatzen du.

- **Itzulpenen doitasuna:** Ematen diren itzulpenak zuzenak diren ala ez. Maila ezberdinetako zuzentasun mailak definitu daitezke, izan ere, hiztegi horien erabilpen testuinguru batzuetan zorroztasun maila beste batzuetan baino baxuagoa izan daiteke. Adibide bat jartzearren, hizkuntza ezezagun batean idatzia dagoen testu bat ulertzeko, zuzenak izan ez arren antzeko esanahiak dituzten itzulpenak ematen dituen hiztegi bat oso erabilgarria izan daiteke.

- **Itzulpen probableenen estaldura:** Sarrera batek adiera bat baino gehiago duen kasuetan, normalean adiera bat beste guztiak baino arruntagoa izaten da, hau da, kasu gehienetan sarrera hori adiera horrekin erabiltzen da. Ezaugarri honek adiera probableenen estaldura neurtzen du.

- **maiztasun-faktorea:** Hiztegi elebidun bateko sarrerak sailkatzeko beste irizpide bat erabilera maiztasuna izan daiteke, hau da, kasu errealetan zein maiztasunekin

agertzen edo erabiltzen diren sarrera horiek. Izan ere, hiztegi elebidun baten kasuan interesgarriagoa dirudi maizen erabiltzen diren sarrerak eta horien itzulpenak maiztasun txikiagokoak edo arraroagoak direnak baino hobeto tratatzea. Irizpide hori finkaturik, orain arte definitu diren ezaugarri guztiak irizpide honekiko neurtu daitezke:

- *Sarreren estaldura maiztasuna kontuan hartuta*
- *Itzulpenen estaldura maiztasuna kontuan hartuta*
- *Itzulpenen doitasuna maiztasuna kontuan hartuta*
- *Itzulpen probableenen estaldura maiztasuna kontuan hartuta*

Sarreren erabilpen maiztasunak 3.3.1 atalean aurkezturiko euskarazko corpus elebakarretik eskuratu dira, esperimentu guztietan jatorri hizkuntza bezala euskara erabiltzen baita.

- **Kategoria gramatikalaren faktorea:** Hiztegi elebidun bateko sarrerak sailkatzeko beste irizpide bat kategoria gramatikala izan daiteke (POS). Izan ere, interesgarria izan daiteke kategoria zehatz bateko sarrerak beste kategorietakoak baino hobeto edo okerrago tratatzen diren aztertzea.

Ezaugarri horien ebaluazioa egiteko bi aukera nagusi daude: azterketa eskuz egitea edo azterketa modu automatikoan egitea. Aukera bakoitzak bere alde positiboak eta alde negatiboak ditu.

4.1.3.1 Eskuzko azterketa

Emaitzen eskuzko azterketa egin ahal izateko beharrezkoa da gutxienez aztertu nahi den hiztegi elebidunak lantzen dituen bi hizkuntzak ondo ezagutzen dituen ebaluatzaile bat lortzea. Gainera, eta kasu askotan ager daitezkeen subjektibitate kasuak konpontzeko ebaluatzaile bat baino gehiagoren epaiak konparatu ohi dira aztertzen diren itzulpen bikoteak benetan zuzenak diren erabakitzeko. Bestalde, hiztegi osoaren lagin txiki bat bakarrik aztertuta ere, prozesu honek suposatzen duen lan eskerga izugarria da.

Aztertu ahal diren ezaugarriei begiratuz gero, nabarmena da sarreren estaldura edota itzulpenen estaldura bezalako ezaugarriak eskuz ebaluatzea ezinezkoa dela, izan ere, oso zaila da hizkuntza batek erabiltzen duen lexiko guztia ezagutzea, eta are zailagoa lexiko horretako sarrera bakoitzak dituen adiera guztiak ezagutzea. Gainera, ahal izanda ere, suposatuko lukeen lan eskerga ez da bideragarria. Aldiz, itzulpenen doitasuna edota itzulpen probableenen estaldura guztiz fidagarriak dira (ebaluatzaileak hizkuntzak maila egokian ezagutzearen baldintza betetzen bada noski), emaitza eskuz errebisatzen baita. Gainera doitasun maila ezberdinak definitzeko aukera ematen da (itzulpenak zuzenak ez izan arren ulergarriak direla definitzeko adibidez). Beraz, eskuzko azterketa batek emaitza fidagarriak ematen dituen arren lan handia suposatzen du. Gainera lantzen diren bi hizkuntzak ondo ezagutzea baldintza oso gogorra izan daiteke hizkuntza konbinazio konkretu batzuen kasuan (Euskara-Swahili esaterako).

4.1.3.2 Azterketa automatikoa

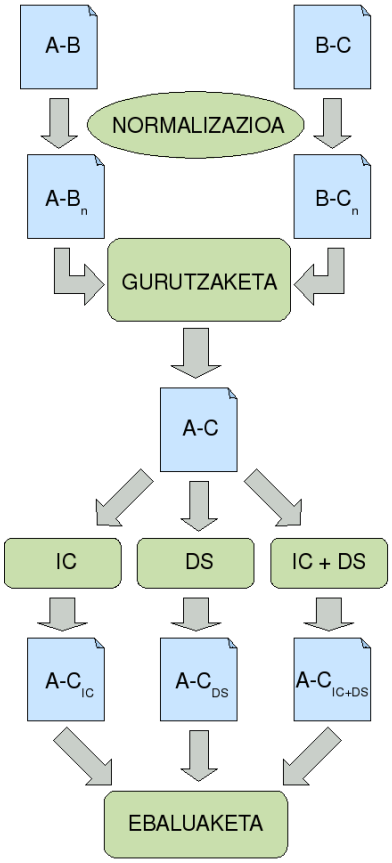
Emaitzen azterketa automatikoa egin ahal izateko beharrezko baldintza da aztertu nahi den hiztegi elebidunak lantzen dituen bi hizkuntzak barneratzen dituen hiztegi elebidun bat eskura izatea. Hiztegi honek erreferentziazko hiztegi bezala funtzionatuko du, eta beraz, hiztegi horretatik kanpo dagoen edozer ez da zuzena izango. Hori dela eta, erreferentzia bezala erabiliko den hiztegi elebidunak lantzen dituen hizkuntzak ondo adierazten dituela bermatu beharra dago. Baldintza hori betetzea oso konplexua da, izan ere hizkuntza batek barneratzen duen lexiko guztia adieraztea ia ezinezkoa izan daiteke, are gehiago bi hizkuntza aldi berean lantzen direnean. Arazo honen ondorioz, zuzenak diren itzulpen bikote asko okertzat har daitezke erreferentzia hiztegian agertzen ez diren aldaerak erabiltzen badira.

Aztertu ahal diren ezaugarriei begiratuz gero, nabarmena da sarreren estaldura edota itzulpenen estaldura nahiko erraz estima daitezkeen arren (erabiltzen den erreferentzia hiztegia onargarria dela suposatuz noski), itzulpenen doitasun estimazioa ez da batere fidagarria erreferentzia hiztegitik kanpo geratzen diren kasu zuzenak okertzat jotzearen ondorioz. Itzulpen probableenen azterketak ere doitasun kalkuluaren arazo berberak planteatzen ditu, eta gainera erreferentzia hiztegiak itzulpenak garrantzia edo erabileraren arabera bereiziak izatera behartzen du. Beraz, azterketa automatikoak eskuzkoak baino lan karga txikiagoa suposatzen badu ere, bere bidez lorturiko emaitzak erreferentzia hiztegiaren fidagarritasunaren menpe daude, fidagarritasun hori estimatzea benetan lan konplexua delarik.

4.2 Esperimentuen diseinua

Aurreko puntuetako faktore guztiak kontuan hartuz, lan honetan bi esperimentu burutzea erabaki da, bi helburu hizkuntza ezberdin erabiliz. Aukeraturiko bi hizkuntzak gaztelera eta txinera dira, eta beraz, eraikiko diren hiztegi elebidunak Euskara-Gaztelera eta Euskara-Txinera izango dira. Konbinazio horiek aukeratzeko aurreko puntuan definituriko faktoreak kontuan hartu dira:

- **Baliabide-faktoreak:** Bi esperimentuak burutzeko beharrezko baliabide multzoa existitzen da (Ikus. 3 kapitulua).
- **Hizkuntzen gertutasuna:** Nabaria da gaztelerak eta txinerak euskararekin duten erlazio oso ezberdina dela: gaztelera eta euskara oso erlazionatuak dauden bitartean, euskara eta txinera urruneko hizkuntzak direla esan daiteke. Horrela, euskararekiko gertutasun maila ezberdina duten hizkuntzak erabiliz, pibotaje tekniken bidez sorturiko hiztegietan faktore honek zenbatean eragiten duen azter daiteke.
- **Emaitzen ebaluazio-prozesua:** Sorturiko hiztegien ezaugarriak modu ezberdinetan aztertzeko helburuarekin, eta ebaluazio mota bakoitzak bere gabeziak dituela kontuan hartuz, sorturiko hiztegi elebidun bat eskuz eta bestea automatikoki aztertzea erabaki da. Horretarako, Euskara-Gaztelera hiztegi elebiduna erreferentzia hiztegi bat erabiliz ebaluatuko da eta Euskara-Txinera hiztegia berriz eskuz.



Irudia 4.1: Esperimentuetan jarraitutako metodologia eskema

Bi esperimentuetan metodologia eta prozesu berberak erabiliko dira, aldatzen den bakarra ebaluazio-prozesua izango delarik. Hurrengo puntuetan esperimentuak burutzeko jarraituriko prozesua (4.1 irudia) sakonago deskribatuko da.

4.2.1 Hiztegien normalizazioa

Esperimentuetan erabiliko diren hiztegiak iturri ezberdinetatik lortu direnez, horietan erabiltzen diren egiturak ezberdinak dira hiztegi bakoitzean (3.2 atala). Hori dela eta, gurutzaketa-prozesua burutu baino lehen beharrezkoa da hiztegi bakoitza formatu eta egitura komun batera esportatzea. Normalizazio fase honen helburua 3.1 atalean aurkeztutako *Apertium* baliabidearen tresnek onartu eta ezagutuko duten XML formatuko hiztegiak lortzea da. Hori baino lehen, hiztegi horien edukiek formatu arau minimo batzuk betetzea bermatu beharko da, gurutzaketa fasea ahalik eta emankorrena izan dadin. Arau horiek hiztegien eduki guztian aplikatzen diren arren, funtsean pibote edo zubi bezala jokatzen duen hizkuntzan dute garrantzirik handiena, izan ere, horiek dira gurutzaketa-prozesuan konparatzen direnak.

HAP masterra

- **Espaziorik ez:** Hitz ugariko sarrerak edo itzulpenak agertzen diren kasuetan, espazioak "␣" ikurragatik ordezkatzen dira.
- **XML karaktere ordezkapena:** Normalizazioaren emaitza XML fitxategi bat denez, emaitza hiztegiaren zuzentasuna bermatzeko beharrezkoa izango da XML formatuak erabiltzen dituen karaktere eta ikur bereziak entitateekin ordeztea (Adibidez: *"<"* → *"<"*).
- **Eduki gordina bakarrik:** Zenbait hiztegitan itzulpenetan informazio gehigarria ager daiteke, itzulpenaren eduki gordina desitxuratuz (Adibidez: *"frijitu: (a piece of) fried food"* ordez, *"frijitu: fried food"*). Kategoria gramatikal edota klase semantikoen informazioa ematen duten etiketak ere ezabatu egiten dira.
- **Hizkuntza bakoitzaren berezitasunak:** Esperimentuetan erabiltzen diren hiztegietan hizkuntza bakoitzaren berezitasunak modu ezberdinetara tratatzen dira (ingelesezko aditzetan *"to"* agertu edo ez, edo gaztelerazko genero bereizketak *"doctor/doctora"*). Esperimentu bakoitzean erabiltzen diren hiztegien arabera irizpide zehatzak finkatzea beharrezkoa izango da.

Hiztegien edukiak normalizatu ondoren testu formatu arruntetik XML formatura esportatu beharra dago. Honekin batera, *Apertium* baliabidearen gurutzaketa tresnak behar dituen baliabide guztiak definitzeko 3.1.1.2 atalean azalduriko hiztegi elebakarrak ere sortu beharra dago. Normalizazio fasearen amaieran hiztegi elebidun bakoitzeko hiztegi elebakar bat eta hiztegi elebidun bat lortzen dira, biak XML formatuan kodetuak eta gurutzaketa-prozesuan erabiltzeko prest.

4.2.2 Gurutzaketa-prozesua

Gurutzatu nahi diren hiztegiak normalizatu eta XML formatura esportatu ondoren, bi hiztegi elebidunen gurutzaketa burutzen da, horretarako 3.1 atalean azaldutako *Apertium* tresnaren hiztegi gurutzaketarako baliabideak erabiliz *(Crossdics)*. Gurutzaketa-prozesuari ahal diren baldintza malguenak esleitzen zaizkio: pibote edo zubi hitz bat edo gehiago komunean dituzten sarrera eta itzulpenak itzulpen bikotetzat hartzen dira. Sarreren kategoria gramatikalak edota klase semantikoak ez dira kontuan hartzen, informazio mota hori ez baita hiztegi guztietan agertzen (are gutxiago baliabide gutxien duten hizkuntzen kasuan). Prozesu honek besteak beste gurutzaketaren emaitza nagusia den hiztegi elebiduna itzultzen du, XML formatuan. Esperimentuen hurrengo faseetan tratamendu lanak errazteko, XML formaturik testu gordinerako esportazio bat burutzen da. Esportazio honek lerro bakoitzean sarrera bat eta bere itzulpen bat ematen ditu. Sarrera batek itzulpen bat baino gehiago duen kasuetan, itzulpen bakoitzeko lerro bat gehitzen da sarrera errepikatuz:

...
etxe hogar
etxe casa
...

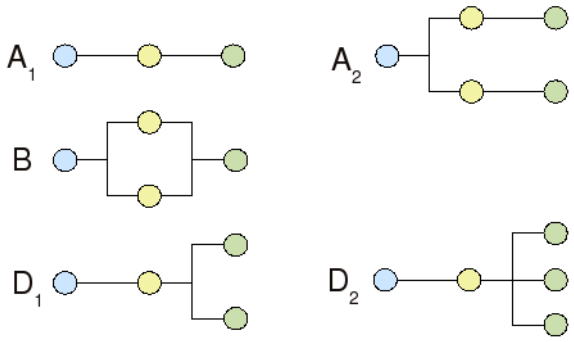
4.2.3 *IC* teknika

Hasierako hiztegi elebidunen bidez sorturiko hiztegi gurutzatu berriei aplikatzen zaien tekniketako bat *IC (Inverse Consultation)* teknika da. 2.2.1.1 puntuan deskribatu bezala, teknika hori gurutzatzen diren hiztegi elebidunen egituran oinarritzen da sarrera eta itzulpenen antzekotasun semantikoa estimatzeko. Horrela, gero eta lotura gehiago aurkitu sarrera eta itzulpen zehatz baten artean, orduan eta antzekotasun semantiko handiagoa dutela suposatzen da. Lan honetan burutzen diren esperimentuetan, Tanaka eta Umemura (1994) lanean sarrera bakoitza kalifikatzeko erabilitako sailkapenen sistema hartzen da oinarritzat, aldaketa minimo batzuk aplikatzen zaizkiolarik. Horrela, guztira 6 sarrera mota desberdin definitzen dira, sarrera horien eta beren itzulpenen artean dagoen lotura egituraren arabera (ikus 4.2 irudia):

- **A_1 mota:** Sarrerak itzulpen bakarra du, pibote edo zubi hitz bakarraren bidez lotua. Bikote horien antzekotasun semantikoa handia da, sarrera eta pibote hitza monosemikoak direla suposatzen baita.
- **A_2 mota:** Sarrerak itzulpen ugari ditu, baina itzulpen bakoitza zubi hitz bakarrarekin lotzen da. Bikote horien antzekotasun semantikoa nahiko handia da, pibote hitzak monosemikoak direla suposatzen baita.
- **B mota:** Sarrerak *IC* teknikaren baldintza betetzen du, hau da, A_1 edo A_2 motako sarreren baldintzak ez ditu betetzen baina sarrera eta bere itzulpenak gutxienez bi pibote hitzen bidez lotzen dira. Mota honetako sarrerak izango dira esperimentuetan aztertuko direnak.
- **D_1 mota:** Sarrerak ez ditu A_1, A_2 eta B motako sarreren baldintzak betetzen, baina gehienez ere bi itzulpen bakarrik ditu. Zalantzazko itzulpen bikoteak kontsideratzen dira, pibote hitza polisemikoa izatearen ondorioz bikote okerrak sortzeko arriskua baitago.
- **D_2 mota:** Sarrerak ez ditu A_1, A_2 eta B motako sarreren baldintzak betetzen, eta gainera bi itzulpen baino gehiago ditu. Itzulpen bikote arriskutsuak kontsideratzen dira, pibote hitza polisemikoa izatearen ondorioz bikote okerrak sortzeko arriskua baitago (D_1 motakoetan bezala baina maila altuagoan).
- **E mota:** Sarrerari ez zaio itzulpenik lotu gurutzaketa-prozesuan.

Sarrera bateko itzulpenek klase horietako bat baino gehiagoko baldintzak betetzen dituzten kasuetan (ohikoenak *B*, D_1 eta D_2 klaseen konbinazioak izan ohi dira), klase altueneko bikoteak bakarrik hartzen dira kontuan eta gainerakoak prozesutik kanpo geratzen dira, sarrera bakoitzaren itzulpen ziurrenak bakarrik hartu nahi baitira kontuan. Adibidez, demagun s_1 sarrerak *B* motako itzulpen bat *(i_1)* eta D_1 motako 2 itzulpen dituela (i_2 eta i_3). Sarrera bakoitzeko itzulpen ziurrenak bakarrik lortu nahi direnez, kasu honetan s_1 sarrera *B* motakoa dela suposatuko litzateke, eta bere itzulpen bezala i_1 itzulpena bakarrik

mantenduko litzateke, i_2 eta i_3 itzulpenak baino ziurragoa baita. Ondorioz, prozesu honek sarrera bakoitzeko hainbat itzulpen baztertzen dituela onartzen da. Gure esperimentuetan B motako sarrerak izango dira aztergaia, horiek baitira IC teknikak emaitza bezala proposatzen dituen sarrerak eta itzulpenak. Bestalde, A_1 eta A_2 motako sarrerak eta horien itzulpenak zuzen bezala hartuko dira automatikoki, irizpideetan aipatu bezala polisemia maila baxuena erakusten duten bikoteak izaten baitira bi mota horietakoak. Hortaz, IC teknika aplikatuz sorturiko hiztegietan A_1, A_2 eta B motako sarrerak agertuko dira (A_1 eta A_2 multzotakoak ziurtzat hartzen direlako eta B motakoak IC teknikak ezarritako baldintzak gainditzen dituztelako).



Irudia 4.2: Sarreren sailkapen eskemaren egitura adibideak

4.2.4 *DS* teknika

Hasierako hiztegi elebidunen bidez sorturiko hiztegi gurutzatu berriei aplikatzen zaien tekniketako bat *DS (Distributional Similarity)* teknika da. 2.2.2.1 puntuan deskribatu bezala, teknika hori itzulpen bikoteetako sarrerek eta itzulpenek corpus konparagarrietan dituzten testuinguruen antzekotasun mailan oinarritzen da horien antzekotasun semantikoa estimatzeko. Horrela, antzeko esanahia duten hitzak beti antzeko testuinguruetan agertuko direla suposatzen da. Lan honetan burutzen diren esperimentuetan, sarrera eta itzulpenen arteko antzekotasun distribuzionala kalkulatzeko hurbilpen tradizionala erabili da (Fung (1995); Rapp (1999)). Horrela, hitz bakoitzaren testuingurua bere aurretik eta ondoren agertzen diren 5 hitzen bidez osatzen da (± 5). Testuingurua osatzen duten hitzek aztertzen den hitzarekiko duten pisua neurtzeko *LLR (Log-likelihood)* neurria aplikatzen da, horrela hitz bakoitzeko bere testuingurua adierazten duen bektorea osatzen delarik. Hitz bakoitzeko bere testuinguru bektorea lortu ostean, sarrera hizkuntzako hitz bakoitzaren testuinguru bektorearen eta helburu hizkuntzako bektore guztien arteko antzekotasuna kalkulatzen da, horretarako bektoreen arteko kosinua kalkulatzen delarik. Hizkuntzarteko antzekotasuna kalkulatu ahal izateko, nabaria da testuinguru bektoreak itzuli beharra dagoela, horretarako hiztegi elebidunak erabili behar direlarik. Kasu honetan, itzulpena egiteko erabili beharko litzatekeen hiztegi elebiduna prozesu osoaren emaitza da, beraz itzulpen horiek egiteko baliabiderik ez dugula suposatu behar da. Honen aurrean, aukera posible bat, gu-

rutzaketa gordinak emaitza bezala itzultzen duen hiztegi zaratatsua erabiltzea da. Hiztegi honetatik hainbat bertsio ezberdin probatuko dira:

- Gurutzaketa osteko hiztegiko sarrera monosemikoak biltzen dituen hiztegia.
- Gurutzaketa osteko hiztegiko sarrera monosemikoak eta sarrera polisemikoen itzulpen maizenak (helburu hizkuntzako corpusaren arabera) biltzen dituen hiztegia.
- *IC* metodoa aplikatuz lortzen den hiztegia.

Gure esperimentuetan, antzekotasun atalase ezberdinak aplikatuko dira eta kasu bakoitzean atalase hori gainditzen duten sarrerak izango dira aztergaia. Atalase hauek lokalak (sarrera bakoitzeko) edo globalak (balio orokorrekiko) izan daitezke. Bestalde, sarrera monosemikoak eta horien itzulpenak zuzen bezala hartuko dira automatikoki, irizpideetan aipatu bezala polisemia maila baxuena erakusten duten bikoteak izaten baitira bi mota horietakoak. Hortaz, *DS* teknikaren bidez sorturiko hiztegietan sarrera monosemikoak eta antzekotasun atalase finkatua gainditzen duten sarrerak agertuko dira.

4.2.5 *IC* eta *DS* tekniken konbinaketa

Aurreko ataletan azaldu bezala, *IC* eta *DS* teknikak dira baliabide gutxien duten hizkuntzekin lan egiteko aukera egokienak, biak nahiko erraz lor daitezkeen baliabideekin lan egiten baitute. Bestalde, bi teknikak ideia edo paradigma eta baliabide ezberdinetan oinarritzen dira. *IC*-ren kasuan, itzulpen bikoteen aukeraketa-prozesuan parte hartzen duten hiztegien egituren menpe dago. *DS*-ren kasuan berriz, aukeraketa-prozesua itzulpen hautagaiek corpusetan dituzten testuinguruen azterketan oinarritzen da. Hori horrela izanik, lan honetan bi teknika horiek konbinatzea proposatzen da, horien emaitzek dibergentzia maila minimo bat izango dutela pentsa baitaiteke. Planteatzen den ideia hau da: *IC* teknikak *DS* teknikak lortzen ez dituen itzulpen bikote batzuk lortzea gerta daiteke, eta alderantziz, *DS* teknikak *IC* teknikak lortu ezin dituen itzulpen bikote batzuk lortzea ere gerta daiteke. Dibergentzia maila horri esker teknikak banaka erabiliz lorturiko emaitzak gainditzerik dagoen aztertu behar da.

IC eta *DS* teknikak hainbat eratara konbina daitezke:

- **Konbinaketa gordina *(UNION)*:** Bi teknikekin lorturiko emaitzak (itzulpen bikoteak) bateratu egiten dira. Itzulpen bikoteen errepikapenak bigarren urrats batean garbitzen dira azken emaitzan bikote bakoitza behin bakarrik agertzeko. *IC* teknikaren kasuan, bikote guztiak hartzen dira kontuan bateratzea egiteko momentuan, aldiz *DS* teknikaren kasuan, atalase zehatz bat gainditzen duten itzulpen bikoteak bakarrik hartzen dira kontuan.
- **Konbinaketa lineala *(LCOMB)*:** Lantzen den itzulpen bikote bakoitzeko, teknika bakoitzak ziurtasun balio bat esleitzen dio: *IC* teknikaren kasuan, balio hori sarrera eta itzulpenaren arteko loturan parte hartzen duen pibote hitz kopurua da, eta *DS* teknikaren kasuan, sarrera eta itzulpenaren arteko antzekotasun distribuzionala

(corpus konparagarrietatik lortua). Balio horiek linealki konbinatzen dira itzulpen bikote hautagai bakoitzari balio bakar bat esleitzeko. Balio konbinatuaren atalase zehatz bat gainditzen duten itzulpen bikote hautagaiak emaitza hiztegian gehitzen dira:

$$LCOMB = IC * k + DS * (1 - k)$$

4.2.6 Emaitzak ebaluatzeko sistema

Puntu honen hasieran aipatu bezala, lan honetan bi esperimentu burutuko dira, bata eskuz ebaluatuko delarik (Euskara-Txinera) eta bestea automatikoki (Euskara-Gaztelera). Bi kasuetan, esperimentuen helburu nagusiak honakoak izango dira:

1. ***IC* eta *DS* tekniken arteko konparaketa bat burutu:** Teknika bakoitzak kasu ezberdinetan duen errendimendua ebaluatuko da lehendabizi eta jarraian bi tekniken emaitzak konparatuko dira. Honen helburua teknika bakoitza zein egoeratan den erabilgarriagoa estimatzea da.

2. ***IC* eta *DS* tekniken konbinazioaren emaitzak aztertu:** *IC* eta *DS* teknikak paradigma eta baliabide ezberdinetan oinarritzen direla jakinik, beraien emaitzen artean dagoen dibergentzia maila zenbatekoa den estimatuko da horrela konbinaketaren bidez lor daitekeen hobekuntza estimatzeko.

3. **Lorturiko emaitza hiztegien kalitatea estimatu:** *IC*, *DS* eta bien konbinaketaren bidez lorturiko hiztegien kalitatea estimatuko da. Horrela, hiztegi horiek erabilpen kasu errealetan izango luketen erabilgarritasuna zenbatekoa den neurtu daiteke.

Hiru helburu nagusi horiek betetzeko, erabilpen kasu edo testuinguru ezberdinak planteatuko dira, esperimentuen emaitzak erabilpen testuinguruaren arabera nola aldatzen diren aztertzeko. Horrela, testuinguru bakoitzean emaitza erabilgarrienak zein teknikak lortu ditzakeen aztertuko da. Erabilpen testuinguru bakoitza metrika ezberdin baten bidez estimatuko da.

4.2.6.1 Euskara-Gaztelera hiztegiaren ebaluazio-sistema

Aurreko puntuetan aurreratu bezala, Euskara-Gaztelera hiztegi elebidunaren ebaluazioa modu automatikoan burutuko da. Horretarako, lehendabizi *IC* teknika, *DS* teknika eta bi tekniken konbinazioa modu isolatuan aplikatuko dira teknika bakoitzaren bidez hiztegi elebidun ezberdinak sortzeko. Hiztegi horietariko bakoitzeko honako propietateak (4.1.3 puntuan definituak) aztertuko dira erreferentzia bezala 3.2.3 puntuan aurkezturiko Euskara-Gaztelera hiztegi elebiduna erabiliz:

- **Sarrera bakoitzeko itzulpenen bataz besteko doitasuna sarreren maiztasunarekiko neurtua.**

HAP masterra

- **Sarrera bakoitzeko itzulpenen bataz besteko estaldura sarreren maiztasunarekiko neurtua.**

- **Sarrera bakoitzeko itzulpenen bataz besteko F-scorea.** Neurri honen bidez bataz besteko doitasuna eta estaldura erlazionatzen dira:

$$F = 2 * \frac{(doitasuna * estaldura)}{(doitasuna + estaldura)}$$

- **Sarrera bakoitzeko itzulpenen bataz besteko $\mathbf{F_2}$.** Neurri honen bidez bataz besteko doitasuna eta estaldura erlazionatzen dira, baina estaldurari garrantzia handiagoa ematen zaio. Horrela, hiztegien tamaina garrantzitsua den testuinguruetan aztertzen diren hiztegiek duten egokitasuna ebaluatzen da:

$$F_2 = (1 + 2^2) * \frac{(doitasuna * estaldura)}{(2^2 * doitasuna + estaldura)}$$

- **Sarrera bakoitzeko itzulpenen bataz besteko $\mathbf{F_{0,5}}$.** Neurri honen bidez bataz besteko doitasuna eta estaldura erlazionatzen dira, baina doitasunari garrantzia handiagoa ematen zaio. Horrela, hiztegien zuzentasuna garrantzitsua den testuinguruetan aztertzen diren hiztegiek duten egokitasuna ebaluatzen da:

$$F_{0,5} = (1 + 0,5^2) * \frac{(doitasuna * estaldura)}{(0,5^2 * doitasuna + estaldura)}$$

- **Sarrera bakoitzeko itzulpen probableenen estaldura.** Neurri honen bidez sarrera bakoitzeko itzulpenen izaera aztertzen da, hau da, itzulpenak komunenak diren (ohikoenak) edota aldiz arraroagoak diren (ez hain erabiliak edo ez ohikoak). Sarrera bakoitzeko ohikoenak diren itzulpenak identifikazioa 3.3.4.1 atalean azalduriko *EU-ES* corpus paraleloan oinarritzen da. Horretarako, landu nahi den sarrera bakoitzeko bere itzulpen bikote guztiak eta horien agerpen maiztasunak erauzten dira corpus paralelotik. Sarrera bakoitzeko maizen erabiltzen diren itzulpenak itzulpen probable edo ohikoenak direla suposa daiteke. Ideia honetan oinarrituz, esperimentuetan sortutako *EU-ES* hiztegi elebidunek itzultzen dituen itzulpenak ohikoenak diren ala ez ebaluatu daiteke.

Teknika bakoitzeko propietate horien azterketa egin ostean, teknika guztien bidez lorturiko emaitzak konparatuko dira horrela teknika horietariko bakoitza zein egoeratan den erabilgarriagoa ondorioztatu ahal izateko.

4.2.6.2 Euskara-Txinera hiztegiaren ebaluazio-sistema

Aurreko puntuetan aurreratu bezala, Euskara-Txinera hiztegi elebidunaren ebaluazioa eskuz burutuko da. Horretarako lehendabizi *IC* teknika eta *DS* teknika aplikatuko dira. Jarraian, teknika bakoitzarekin lorturiko itzulpen bikote guztien eta gurutzaketaren ostean lortzen diren itzulpen bikote monosemikoen bidez hiztegi konbinatu bat sortuko da. Esperimentuaren helburua hiztegi honen azterketa izango da, horrela tekniken portaera nahiz lorturiko hiztegiaren kalitatea neurtzeko. Hiztegi osoaren eskuzko azterketa egiteak lan eskerga handiegia suposatzen duenez, hiztegi honen lagin bat aukeratuko da hausaz. Lagina sortzeko aplikatuko diren irizpideak hurrengo puntuan laburtzen dira.

4.2.6.2.1 Euskara-Txinera ebaluazio lagina

EU-ZH hiztegi lagina sortzeko aplikatzen diren irizpideak honakoak dira:

- Esperimentuaren helburuetako bat *IC* teknika eta *DS* teknika aztertzea denez, bi teknikak kondizio beretan ebaluatu ahal izateko bi tekniken baldintzak gainditzen dituzten sarrerak bakarrik aztertuko dira.
- Maiztasun tarte guztietako sarrerak aztertu ahal izateko, 3 maiztasun tarte aplikatuko dira eta tarte bakoitzeko sarrera kopuru finko bat gehituko da laginean (hausaz aukeratuak).
- Aukeratzen den laginaren ezaugarriak hiztegi osoaren ahalik eta antzekoenak izatea komenigarria da. Horretarako aukeratzen diren sarrerek ondoko baldintzak betetzea bermatuko da: Sarrera polisemiko/monosemikoen arteko oreka mantenduko da, beraz, lagina sarrera monosemiko eta polisemikoen bidez osatuko da. Multzo bakoitzeko aukeratzen den sarrera kopurua multzo bakoitzak hiztegi osoan duen presentziarekiko proportzionala izango da.
- Aukeraturiko sarrera bakoitzeko bi tekniken bidez lorturiko itzulpen guztiak sartuko dira laginean, nahiz eta agian itzulpen horietako batzuk bi tekniken bidez lortuak ez izan (adibidez, sarrera konkretu batentzat *IC* teknikak 3 itzulpen posible itzultzen baditu eta *DS* teknikak 3 horiek eta gainera beste 5 berri, laginean 8 itzulpen posibleak sartuko dira). Metodologia hori jarraituko da, aurreko hiru puntuetan finkaturiko irizpideak hiztegiko sarreren aukeraketan eragiten dutelako, baina ez sarrera horien itzulpenetan.

4.2.6.2.2 Ebaluazio lagineko sarrerak eta itzulpenak ebaluatzeko kategoria-sistema

Laginaren ebaluazio egin ahal izateko, hainbat epailek hartuko dute parte prozesuan. Epaile horiek itzulpen bakoitzaren zuzentasuna adierazteko ondoko kategoria-sistema erabiliko dute:

HAP masterra

- **Okerra:** Epaileak ziurtasun osoz erabaki du itzulpena okerra dela.
- **Zuzena:** Epaileak ziurtasun osoz erabaki du itzulpena zuzena dela.
- **Kategoria okerra:** Sarrerak eta itzulpenak antzeko esanahia dute baina kategoria gramatikal ezberdina (Adibidez, *"erorketa"* eta *"erori"*).
- **Zalantza:** Eskura duen informazioa erabiliz epailea ez da gai itzulpena zuzena, okerra edo kategoria okerrekoa den erabakitzeko:
  - Antzeko esanahiak dituzte baina baten esanahia bestea baino zabalagoa da.
  - Epaileak laguntza bezala erabiltzen duen hiztegiko informazioa ez da nahikoa erabaki bat hartzeko.

Bestalde, lagina osatzen duten itzulpen bikoteak epaileen artean nola banatuko diren erabaki behar da. Ebaluatzaile taldean laginak barneratzen dituen bi hizkuntzetako hiztunik ez dagoenez, epaien fidagarritasuna handitzeko itzulpen bikote bakoitzak epai bat baino gehiago jasotzea erabaki da. Horrela, itzulpen bikote bakoitzak bi epai jasoko ditu, bi ebaluatzaile ezberdinek emanak. Jasotako epaien arabera itzulpen bikotea laginean sartu edo ez erabakitzeko irizpideak ondokoak izango dira:

- **zuzena/zuzena:** Itzulpen bikotea laginean sartuko da, zuzen bezala.
- **okerra/okerra:** Itzulpen bikotea laginean sartuko da, oker bezala.
- **kategoria okerra/kategoria okerra:** Itzulpen bikotea laginean sartuko da, zuzen bezala.
- **zalantza/zalantza:** Itzulpen bikotea laginetik kanpo geratuko da.
- **zuzena/okerra:** Itzulpen bikotearen epaia errebisatuko da, epai komun bat lortuko delarik.
- **zuzena/kategoria okerra:** Itzulpen bikotea laginean sartuko da, zuzen bezala.
- **zuzena/zalantza:** Itzulpen bikotea laginetik kanpo geratuko da.
- **okerra/kategoria okerra:** Itzulpen bikotea laginetik kanpo geratuko da.
- **okerra/zalantza:** Itzulpen bikotea laginetik kanpo gertuko da.
- **kategoria okerra/zalantza:** Itzulpen bikotea laginetik kanpo geratuko da.

Horrela, ebaluazio lagina goian finkaturiko irizpideak gainditzen dituzten bikoteen bidez osatuko da. Nabaria da hainbat kasutan modu malguan jokatzen dela (batez ere kategoria okerra duten bikoteak zuzentzat hartzen diren kasuetan). Malgutasun honen helburua ahalik eta lagin handiena lortzea da.

HAP masterra

4.2.6.2.3 Euskara-Txinera laginaren ebaluazio-sistema

Euskara-Txinera hiztegi osoaren propietateak estimatzeko erabiltzen den laginaren sarrera bakoitzeko, hainbat ezaugarriak neurtuko dira. Laginaren bidez ateratzen diren ondorio guztiak hiztegi osoari aplikatu ahal zaizkiola suposatuko da, 4.3.6.2.1 puntuan zehazturiko irizpideetan oinarrituz. Horiek dira laginaren bidez aztertuko diren hiztegiaren propietateak:

- **Itzulpenen bataz besteko doitasuna, sarreraren maiztasunarekiko neurtua.**
- **Itzulpenen bataz besteko estaldura, sarreraren maiztasunarekiko neurtua.**
- **Itzulpenen bataz besteko F-scorea.** Neurri honen bidez bataz besteko doitasuna eta estaldura erlazionatzen dira:

$$F = 2 * \frac{(doitasuna * estaldura)}{(doitasuna + estaldura)}$$

- **Itzulpenen bataz besteko $\mathbf{F_2}$.** Neurri honen bidez bataz besteko doitasuna eta estaldura erlazionatzen dira, baina estaldurari garrantzia handiagoa ematen zaio. Horrela, hiztegien tamaina garrantzitsua den testuinguruetan aztertzen diren hiztegiek duten egokitasuna ebaluatzen da:

$$F_2 = (1 + 2^2) * \frac{(doitasuna * estaldura)}{(2^2 * doitasuna + estaldura)}$$

- **Itzulpenen bataz besteko $\mathbf{F_{0,5}}$.** Neurri honen bidez bataz besteko doitasuna eta estaldura erlazionatzen dira, baina doitasunari garrantzia handiagoa ematen zaio. Horrela, hiztegien zuzentasuna garrantzitsua den testuinguruetan aztertzen diren hiztegiek duten egokitasuna ebaluatzen da:

$$F_{0,5} = (1 + 0,5^2) * \frac{(doitasuna * estaldura)}{(0,5^2 * doitasuna + estaldura)}$$

- **Itzulpenen bataz besteko F-scorea sarreren kategoria gramatikalarekiko *(POS)*.** Neurri honen bidez bataz besteko doitasuna eta estaldura erlazionatzen dira, baina emaitzak sarreren kategoria gramatikalaren arabera taldekatzen dira. Beraz, neurri honen helburua kategoria gramatikalaren araberako emaitzen aldakortasuna aztertzea da.

- **Sarrera bakoitzeko itzulpen probableenen estaldura.** Neurri honen bidez sarrera bakoitzeko itzulpenen izaera aztertzen da, hau da, itzulpenak komunenak diren (ohikoenak) edota aldiz arraroagoak diren (ez hain erabiliak edo ez ohikoak). Sarrera bakoitzeko ohikoenak diren itzulpenak identifikazioa 3.3.4.2 atalean azalduriko *EU-ZH* corpus paraleloan oinarritzen da. Horretarako, landu nahi den sarrera bakoitzeko bere itzulpen bikote guztiak eta horien agerpen maiztasunak erauzten dira corpus paralelotik. Sarrera bakoitzeko maizen erabiltzen diren itzulpenak itzulpen probable edo ohikoenak direla suposa daiteke. Ideia honetan oinarrituz, esperimentuetan sortutako *EU-ZH* hiztegi elebidunek itzultzen dituen itzulpenak ohikoenak diren ala ez ebaluatu daiteke.

Teknika bakoitzeko propietate horien azterketa egin ostean, teknika guztien bidez lorturiko emaitzak konparatuko dira horrela teknika horietariko bakoitza zein egoeratan den erabilgarriagoa ondorioztatu ahal izateko.

Bestalde, eskuzko ebaluaziotik kanpo, emaitza hiztegiek gurutzaketa-prozesuan erabiltzen diren hiztegiekiko (*EU-EN* eta *EN-ZH* hiztegiekiko) duten sarrera eta itzulpen estaldurak neurtuko dira. Neurri honen bidez ez da itzulpen bikoteen zuzentasunaren informaziorik kontuan hartzen, kasurik onenean sortzen den hiztegiak gurutzaketan parte hartu duten jatorrizko hiztegietan agertzen diren sarrera eta itzulpen ezberdin guztiak tratatzea zenbateko proportzioan lortzen duen baizik. Azken finean, gurutzaketa-prozesuan parte hartzen duten sarrera eta itzulpen posibleen multzo osotik emaitza hiztegiak tratatzen duen ehunekoa da estimatzen dena. Balio horiek sarreren bataz besteko estaldura eta itzulpenen bataz besteko estaldura bezala definitzen dira eta modu automatikoan neurtuko dira, aztertzen diren itzulpen bikoteen zuzentasunak ez baitu eraginik neurri horien emaitzetan.

5 Emaitzak

Atal honetan zehar aurreko atalean definituriko bi esperimentu handien emaitza guztiak emango dira eta jarraian horietatik atera daitezkeen ondorio nagusiak aztertuko dira. Horretarako, lehendabizi esperimentu bakoitzaren emaitzak modu isolatuan emango dira, jarraian bi esperimentuetan ateratako emaitzetatik ondorio orokorragoak atera ahal izateko. Jarraian, 5.1 eta 5.2 puntuetan lan honetan egindako esperimentuen emaitzak azalduko dira.

5.1 Euskara-Gaztelera hiztegia

Puntu honetan Euskara-Gaztelera hiztegi elebidunaren sorkuntza automatikoak emandako emaitza guztiak azalduko dira. Horretarako, esperimentuaren fase bakoitzeko emaitzak puntu ezberdinetan banatuko dira: lehendabizi gurutzaketatik sorturiko hiztegi gordinaren datuak emango dira eta jarraian teknika bakoitzaren portaera eta horien bidez lorturiko emaitzak azalduko dira (*IC*, *DS* eta *IC-DS* konbinaketa). Bukatzeko, azken puntuetan tekniken konparaketa bat eta esperimentuaren ondorio orokor batzuk azalduko dira.

5.1.1 Gurutzaketa-prozesua

Euskara-Gaztelera hiztegi elebiduna automatikoki sortzeko prozesuan 3.2 atalean definituriko *EU-EN* eta *EN-ES* hiztegiak erabili dira. Horretarako, lehendabizi hiztegi horien edukiak normalizazio-prozesutik pasatu dira bi hiztegien edukien formatuak ahalik eta antzekoenak bihurtzeko. 4.2.1 atalean azaldu bezala, prozesu honen helburua gurutzaketaren emaitzak maximizatzea da. *EU-EN* eta *EN-ES* hiztegi normalizatuen ezaugarri orokorrak eta gurutzaketaren bidez lorturiko *EU-ES* hiztegi berriaren propietate orokorrak 5.1 taulan ikus daitezke.

Hiztegia	Sarrera kopurua	Itzulpen kopurua	Anbiguotasun maila
EU-EN (jatorrizkoa)	17.672	43.021	2,43
EN-ES (jatorrizkoa)	16.326	38.128	2,33
EU-ES (gurutzaketa)	14.012	104.172	7,13

Taula 5.1: *EU-ES* hiztegiaren gurutzaketa-prozesuko hiztegien propietate orokorrak

5.1. taulako datuetan ikus daitekeen bezala, hiztegi gurutzatuaren sarrera kopurua jatorrizko hiztegienaren antzekoa den arren, itzulpen kopurua askoz ere altuagoa da (jatorrizko hiztegien itzulpen kopuruaren bikoitza baino gehiago). Faktore honek eragin zuzena du hiztegiaren anbiguotasun mailan (sarrera bakoitzak bataz beste duen itzulpen kopuruan), izan ere, jatorrizko hiztegien balioa 2 inguru den bitartean gurutzaketaren bidez lorturiko hiztegiarena 7,13-koa da (hau da, sarrera bakoitzeko bataz beste 7,13 itzulpen ematen dira). Hiztegiaren analisi sakonago bat eginez, sarreren %71-ak (10.000 sarrera, 99.844 bikote) itzulpen posible bat baino gehiago du, hau da, polisemikoak dira (gainerakoa %29-a itzulpen bakarreko sarrera monosemikoek osatzen dute). 3.2.2 atalean deskribaturiko *EU-ES*

erreferentzia hiztegia erabiliz egindako azterketa batek, 10.000 sarrera anbiguo horietatik %80,32-ak itzulpen hautagai okerrak dituela erakusten du, hau da, 99.844 bikotetik 80.200 okerrak dira. Datu horiek gurutzaketa-prozesuan parte hartzen duten hiztegien polisemia mailak sorturiko hiztegietan duen eragin kaltegarria estimatzeko balio dute, eta aldi berean gurutzaketaren emaitzak kimatzeko teknika lagungarrien beharra frogatzen da.

Hurrengo puntuetan aplikatzen diren teknikak sarrera polisemikoei bakarrik aplikatuko zaizkie, monosemikoak diren sarrera guztiak ontzat hartuko direlarik. Irizpide hori finkatzeko hipotesia honakoa da: itzulpen bakarra lortu duten sarrerei ez die polisemiak eragin, beraz pentsa daiteke sarrera horien itzulpenak zuzenak izango direla. Hipotesi hori ziurtatzeko, sarrera multzo honetako lagin txiki baten (50 sarrera) gainean egindako eskuzko analisi batek bikote horien doitasuna %100-koa dela frogatu du. Beraz, esperimentu guztian zehar sarrera monosemiko horiek (4.715 sarrera guztira) eta beren itzulpenak emaitzaren parte izango dira eta beti ontzat hartuko dira.

5.1.2 *IC* teknika

Aurreko puntuan sortutako hiztegi gurutzatua eta jatorrizko bi hiztegi elebidunak erabiliz (*EU-EN* hiztegia eta *EN-ES* hiztegia) *IC* teknika aplikatu da. 5.2 taulan 4.2.3 puntuan azaldutako irizpideak jarraituz definituriko sarrera motak eta kopuruak ikus daitezke.

Mota	Sar. kop.	Sar. kop. % totalarekiko	Itz. kopurua	Itz. kop. % totalarekiko
A_1	4.224	%30,14	4.224	%9,09
A_2	491	%3,50	1.066	%2,29
B	3.810	%27,19	7.313	%15,75
D_1	1.415	%10,09	2.830	%6,09
D_2	4.072	%29,06	30.993	%66,75
E	3.657	-	-	-

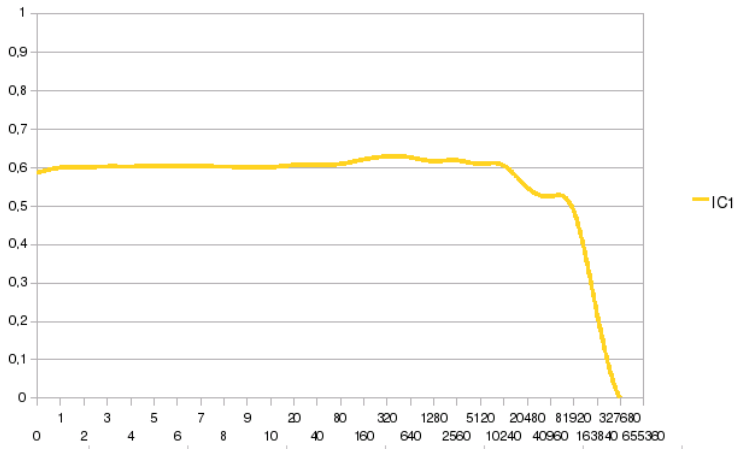
Taula 5.2: Sarrera mota bakoitzeko kopuruak *IC* teknika aplikatu ondoren

Guztira, itzulpen bikoteak dituzten sarrera mota guztietako sarrerak kontuan hartuta (A_1, A_2, B, D_1 eta D_2), 14.012 sarrera eta 46.426 itzulpeneko lagina lortu da. Azpimarratu beharra dago gurutzaketaren ostean lorturiko hiztegiaren itzulpen kopurua (104.172) lagin honetan lortzen dena baina askoz ere handiagoa dela (46.426). Falta diren gainerako itzulpen guztiak (57.746) *IC* teknika aplikatzearen ondorioz galtzen dira, izan ere, eta 4.2.3 puntuan azaldu bezala, sarrera bakoitzeko gutxienez B motako itzulpen bat topatzen denean, sarrera horren D_1 eta D_2 motako itzulpen hautagai guztiak prozesutik kanpo geratzen dira, sailkatzen den elementua sarrera bera baita. Lagin horretatik *IC* teknika gainditzeko baldintzak bete dituzten sarrera bakarrak B motakoak dira. Hortaz, bikote hautagai guztietatik *IC* teknikak finkatzen dituen baldintzak betetzen dituzten sarrerak 3.810 dira (7.313 itzulpen bikote), hau da, gurutzaketa egin ostean lorturiko hiztegiko sarreren %27,1a bakarrik (bikoteen %7-a). Datu horien arabera, *IC* teknika gainditzeko

finkatzen diren baldintzak nahiko gogorrak direla esan daiteke, edo beste era batera esanda, *IC* teknikak bikote hautagai kopuru osoaren zati bat bakarrik tratatu dezake.

IC teknikaren baldintzak gainditzen dituzten sarreren kalitatea neurtu ahal izateko, *B* motako sarrerak eta horien itzulpenak aztertu dira erreferentzia hiztegia erabiliz. Bestalde, *B* motako sarrerei A_1 eta A_2 motako sarrerak (sarrera monosemikoak) ere gehitu zaizkie esperimentua burutzerakoan, izan ere, gogoratu bi sarrera mota horiek beti zuzentzat hartzen direla lan honetan. Erabaki hori hartzearen arrazoi nagusia ahalik eta lagin handienarekin lan egin ahal izatea da.

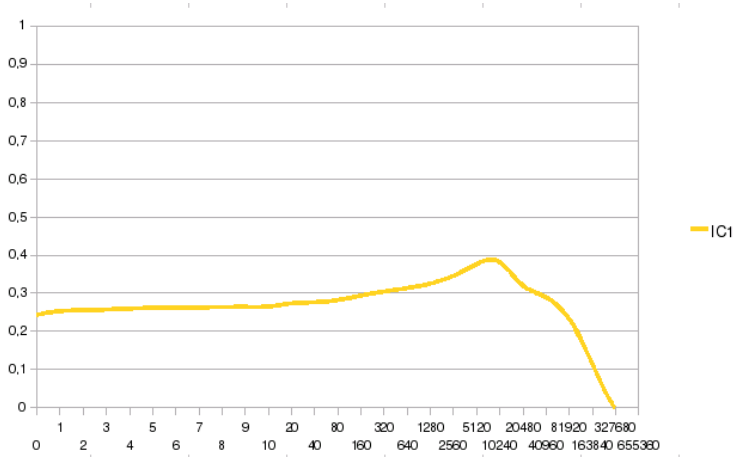
Sarrera bakoitzaren erabilpen maiztasunak *IC* teknikaren portaeran eta lortu den hiztegi elebidunaren izaeran duen eragina aztertu ahal izateko, doitasun eta estaldura balio guztiak sarreren erabilpen maiztasunaren arabera neurtu dira. 5.1 irudian *IC* teknika gainditzen duten sarreren bataz besteko doitasuna ikus daiteke, maiztasunaren arabera.



Irudia 5.1: *IC* teknikaren bidez lorturiko sarreren bataz besteko doitasuna maiztasunarekiko

Oro har, *IC* teknikak lortzen duen doitasuna konstante mantentzen maiztasun txiki eta ertaineko sarreren kasuan ($1<f\leq10.240$), beti ere 0,6 inguruan mantentzen delarik. Maiztasun tarte horretan, maiztasunak gora egin ahala doitasunak ere zertxobait gora egiten du (doitasun maximoa 0,62-ra heltzen da), beraz esan daiteke doitasuna hobea dela maiztasunak gora egiten duen heinean. Maiztasun altueneko sarreren kasuan aldiz ($f>10.240$), doitasunean beherakada handia gertatzen da. Dena den, beherakada hori ez da hain adierazgarria, izan ere maiztasun altueneko sarrera horien kopurua ez da adierazgarria (234 sarrera). Emaitza horien arabera, itzulpen bikote oker askok *IC* teknikaren baldintzak gainditzen dituztela nabarmena da.

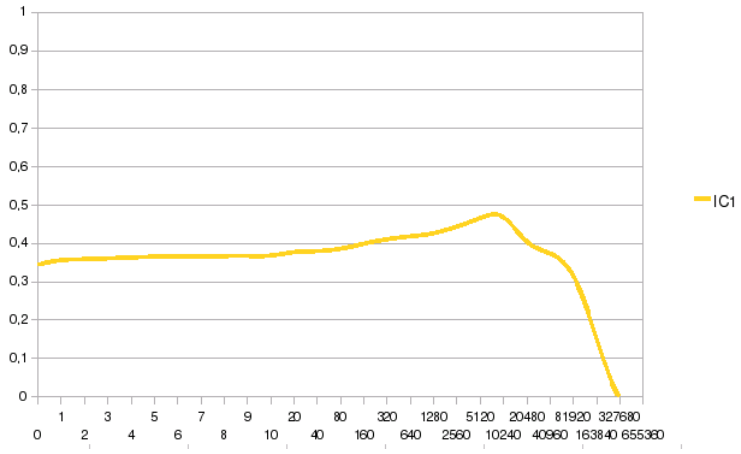
IC teknikaren baldintzak gainditzen dituzten sarreren bataz besteko estalduraren informazioa 5.2 irudian ikus daiteke, maiztasunaren arabera erakutsia.



Irudia 5.2: *IC* teknikaren bidez lorturiko sarreren bataz besteko estaldura maiztasunarekiko

IC teknikak eskaintzen duen bataz besteko estalduraren kasuan nabarmena da orokorrean nahiko baxua dela. Edonola ere, doitasunaren kasuan ikus daitezkeen joera orokorrak mantentzen dira: maiztasun txiki eta ertainetan ($1<f\leq10.240$) estaldurak goraka egiten du, eta maiztasun altuena duten sarreren kasuan ($f>10.240$) aldiz beherakada nabarmena gertatzen da. Beherakada honen zergatia maiztasun altueneko sarreren polisemia maila altua izan daiteke, izan ere, jakina da sarrera maizenek adiera gehiago izateko joera erakusten dutela. Adiera gehiago izateak itzulpen aldaera gehiago edukitzea suposatzen du, eta ondorioz itzulpen guztiak identifikatzea prozesu konplexuagoa bihurtzen da *IC* teknikarentzat. Orokorrean, *IC* teknikak eskaintzen duen estaldura nahiko kaxkarra dela esan daiteke, ziur asko finkatzen diren baldintzen gogortasunaren ondorioz.

Doitasuna eta estalduraren arteko erlazioa definitzen duen F-score neurriaren emaitzak 5.3 irudian ikus daitezke, berriro ere maiztasunaren arabera banatuak.

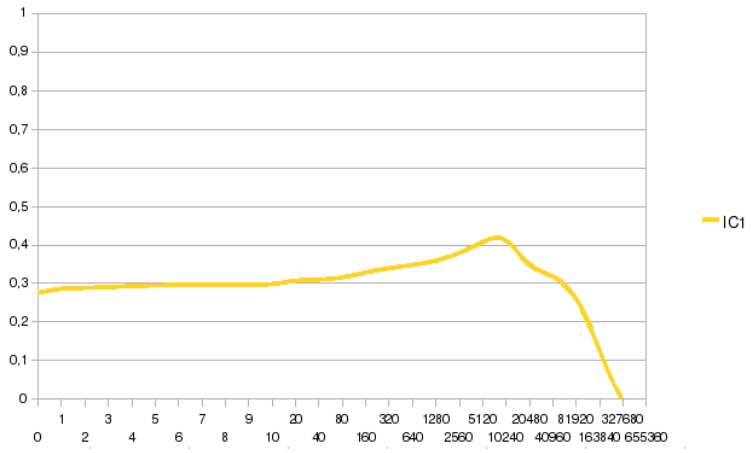


Irudia 5.3: *IC* teknikaren bidez lorturiko sarreren bataz besteko F-scorea maiztasunarekiko

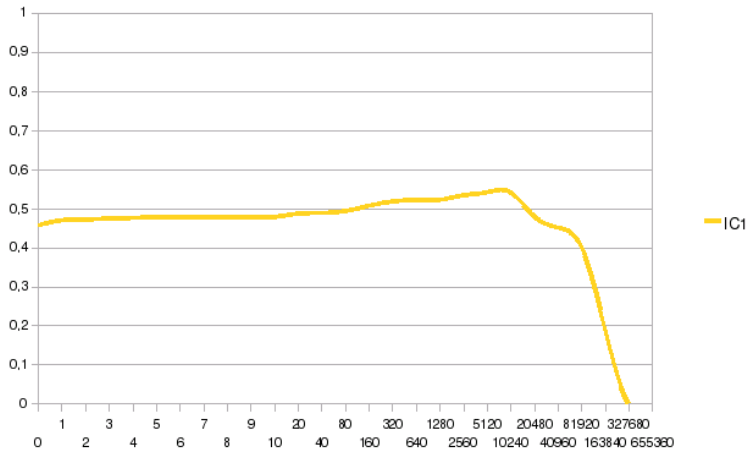
Doitasuna eta estaldura erlazionatzen dituen neurri bat izanik, eta *IC* teknikaren ka-

suan bi neurri horiek sarreren maiztasunarekiko joera berberak jarraitzen dituztela ikusita, F-score neurriaren emaitzek ere joera berberak erakusten dituzte, hau da, maiztasun txiki eta ertaineko sarreren kasuan ($1<f\leq 10.240$) gorakada gertatzen da maiztasunak gora egin ahala, baina maiztasun altuenen kasuan ($f>10.240$) F-scorearen balioak beherakada nabarmena jasaten du.

Bestalde, 4. atalean azaldu bezala, interesgarria da *IC* teknikaren bidez sortzen diren hiztegien emaitzak testuinguru ezberdinetan izan dezaketen erabilgarritasuna estimatzea. Horretarako, F-score neurriaren bi aldaera neurtu dira: estaldurari lehentasuna ematen dion bat (F_2) eta doitasunari garrantzia ematen dion beste bat ($F_{0,5}$). Bi neurri horien emaitzak 5.4 eta 5.5 irudietan ikus daitezke.



Irudia 5.4: *IC* teknikaren bidez lorturiko sarreren bataz besteko F_2 maiztasunarekiko

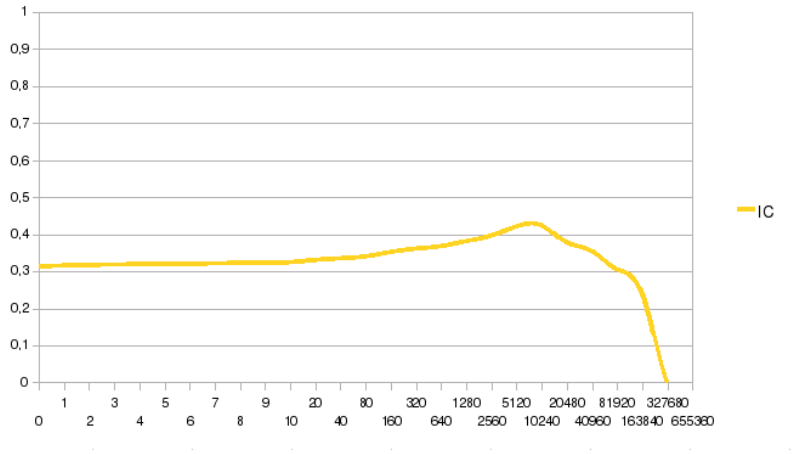


Irudia 5.5: *IC* teknikaren bidez lorturiko sarreren bataz besteko $F_{0,5}$ maiztasunarekiko

F-score neurriaren kasuan gertatzen den bezala, sarreren maiztasunaren araberako joera orokorra mantendu egiten da bietan: maiztasun txiki eta ertainetan neurriek goraka egiten

dute baina maiztasun altueneko sarreren kasuan beherakada gertatzen da. *IC* teknikaren bidez lortzen diren hiztegien doitasunak estaldurak baino balio altuagoak lortzen dituenez, $F_{0,5}$ neurriak F_2 neurriak baino emaitza hobeak lortzen ditu, esan bezala maiztasunaren araberako joerak mantentzen badira ere.

IC teknikaren bidez lortutako sarreretan ebaluatu daitekeen beste alderdi bat ematen dituen itzulpenen izaera da, hau da, ematen dituen itzulpenak komunenak diren (ohikoenak) edota aldiz arraroagoak diren (ez hain erabiliak edo ez ohikoak). 4.2.6.1 atalean azaldu bezala, ebaluazio hori burutu ahal izateko gurutzaketa-prozesuaren ostean sorturiko hiztegi elebiduneko itzulpen bikoteek corpus paralelo batean duten agerpenean oinarritzen gara, corpusean gehien agertzen diren itzulpenak komunenak direla pentsa baitaiteke. Ebaluaziorako erabiliko den erreferentziaren abiapuntua gurutzaketatik sorturiko hiztegi zaratatsua denez, kasu honetan ebaluatu daitekeen faktore bakarra estaldura da (sarrera bakoitzeko itzulpen probableenetatik zenbat ematen diren), izan ere, gurutzaketaren bidez lorturiko hiztegiko bikote guztiak laginean sartzeak doitasunaren kalkulua zentzugabe bihurtzen du (itzulpen bikote guztiak ontzat emango liratekeelako). 5.6 irudian *IC* teknikaren bidez lorturiko sarreren itzulpen problabeenen estaldura-emaitzak ikus daitezke maiztasunaren arabera banatuak.



Irudia 5.6: *IC* teknikaren bidez lorturiko sarreren itzulpen probableenen estaldura maiztasunarekiko

5.6 irudiak erakusten duen moduan, itzulpen probableenen estaldurak gainerako neurriek jarraitzen duten joera berbera erakusten du: maiztasun txiki eta ertaineko sarreren kasuan maiztasuna igo ahala estaldura ere hobea da. Aldiz, maiztasun altueneko sarreren kasuan itzulpen probableenen estaldurak behera egiten du. Oro har, lortzen den estaldura ez da 0,5-era iristen. Ondorioz, esan daiteke *IC* teknikak kasu askotan probableenak ez diren itzulpenak itzultzen dituela (Adibide bat jartzearren, *"usain"* sarreraren itzulpen ohikoena den *"olor"* itzuli beharrean, *"hedor"*, *"peste"* eta *"pestilencia"* itzulpen ez hain ohikoak itzultzen ditu).

Beraz, esan daiteke *IC* teknika batez ere itzulpenen doitasuna garrantzitsua den testuinguruetan bakarrik dela erabilgarria, izan ere lortzen dituen estaldura-emaitzak nahiko

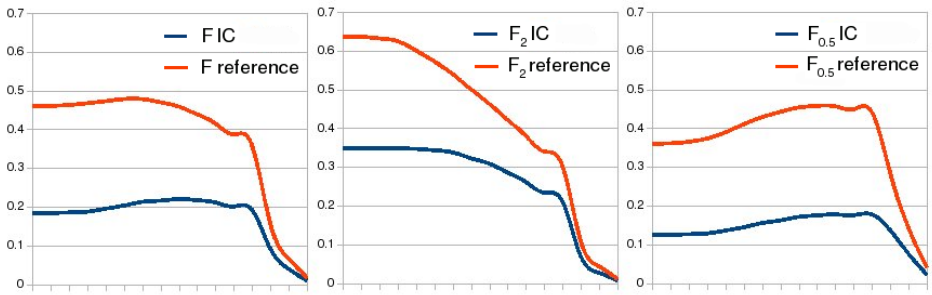
kaxkarrak dira. Emaitza horien kausa nagusia teknikak finkatzen dituen baldintza gogorrak dira, izan ere, gurutzaketa-prozesuan parte hartzen duten bi hiztegi elebidunek sarrera bakoitzeko itzulpen kopuru handia eskaintzea exijitzen du, horrela gurutzaketa-prozesuan ahalik eta pibote edo zubi hitz gehienek parte hartzeko. Sarrera hiztegiek sarrera bakoitzeko itzulpen gutxi ematen dituzten kasuetan, *IC* teknika aplikatu ahal izateko baldintzak gogorregiak bihurtzen dira eta itzulpen bikote askok ez dituzte gainditzen. Aldiz, sarrera bakoitzeko itzulpen asko eskaintzen diren kasuetan, *IC* teknikaren bidez sortzen diren itzulpen bikoteak oso sendoak izaten dira, loturak ebidentzia gehiagotan oinarritzen baitira. Horixe da hain zuzen ere doitasun-emaitza onak lortzearen arrazoi nagusia. Bestalde, sarreren maiztasun-faktoreak teknikaren emaitzetan duen eragina minimoa da, teknikaren portaera nahiko konstante mantentzen delarik maiztasunarekiko (maiztasun altuenetan emaitza okerragoak dira, batez ere sarrera mota horien polisemia maila handiaren ondorioz sortzen diren itzulpen bikote guztiak behar bezala tratatzea oso zaila bihurtzen delako). Azkenik, esperimentuetan ikusitakoaren arabera *IC* teknikak itzultzen dituen itzulpenak kasu askotan ez dira itzulpen probableenak. Fenomeno hori ez da harritzekoa, izan ere *IC* teknika aplikatzeko erabiltzen diren irizpideen bidez ezinezkoa da faktore hori kontrolatzea (hiztegien egituretatik ezinezkoa baita itzulpen probableenak identifikatu ahal izateko informaziorik erauztea). Beraz, *IC* teknika batez ere doitasunak garrantzia handia duen testuinguruetara egokitzen da.

5.1.3 *DS* teknika

Aurreko puntuan sortutako hiztegi gurutzatua, jatorrizko bi hiztegi elebidunak eta bi corpus elebakar konparagarriak erabiliz (*EU-EN* hiztegia, *EN-ES* hiztegia, *EU* corpusa eta *ES* corpusa) *DS* teknika aplikatu da. 4.2.4 puntuan azaldu bezala, *DS* teknika aplikatu ahal izateko eman beharreko lehen urratsa testuinguruen itzulpena egiteko erabiliko den hiztegi elebiduna sortu edo aukeratzea da. Aukera guztiekin esperimentatu ostean, anbiguotasunik ez duten sarrerak eta sarrera polisemikoen itzulpen maizenak (helburu hizkuntzako corpusaren arabera) biltzen dituen hiztegia erabiltzea erabaki da, aurreko kapituluan (4.2.4 puntuan) azalduriko aukeren artean emaitza onenak modu honetara lortzen baitira.

DS teknikaren emaitzak aztertu ahal izateko, lehendabizi emaitzak ontzat hartzeko ezarriko den atalasea definitu beharra dago. 4.2.4 puntuan azaldu bezala, lan honetan bi atalase mota erabiliko ditugu, atalase globalak eta atalase lokalak. Atalase horien balio esperimental optimoak aukeratzeko, lan honetan *IC* teknikaren bidez lorturiko sarrerak eta horien itzulpenak erabiltzea proposatzen da, hau da, itzulpen bikote guztiei aplikatuko zaien atalasea B motako sarreren bidez doitzea. Horretarako, B motako sarreren eta horien itzulpenen arteko antzekotasun distribuzionalak kalkulatzea proposatzen da, eta jarraian F-score maximoa lortzen duten atalase lokalak eta globalak zein diren identifikatzea. Horrela B motako sarrerek osatzen duten laginaren emaitzak hiztegi gurutzatu osoaren emaitzetara orokortzea da helburua. Teknika hori benetan erabilgarria dela frogatzeko, 5.7 irudian *IC* teknikaren bidez lorturiko emaitzen F-scorea eta gurutzaketa osoaren F-scorea ageri dira (F, $F_{0,5}$ eta F_2), atalasearen arabera banatuak (*DS* teknika aplikatu ondoren). Argitu beharra dago esperimentu txiki honen helburu bakarra atalasea finkatzeko erabilitako teknikaren

erabilgarritasuna frogatzea dela.



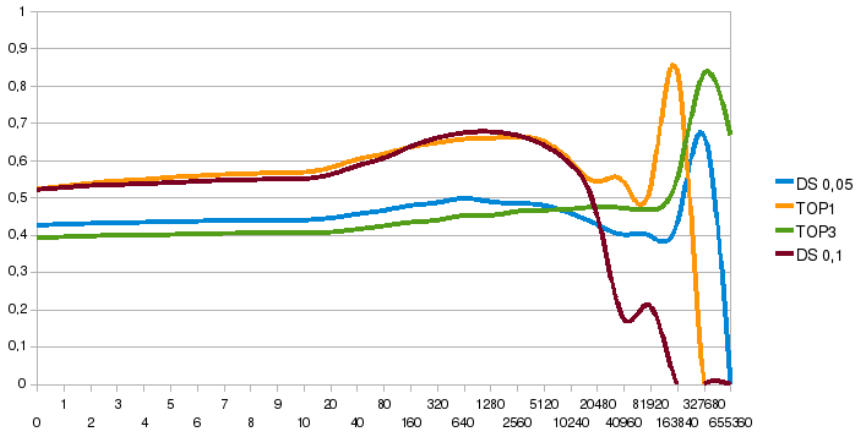
Irudia 5.7: *DS* atalase optimoaren doikuntza: Erreferentzia hiztegia eta *IC* hiztegiaren bidez lorturiko F-scorearen konparaketa

5.7 irudian ikus daitekeenez, *IC* teknikaren bidez lorturiko sarreren emaitza optimoak eta gurutzaketa osoaren bidez lortu daitezkeen emaitza optimoak atalase tarte berdinetan lortzen dira. Teknika hori erabiliz esperimentu honetan erabiliko diren atalase lokal eta globalen balioak erabaki dira. Atalase mota bakoitzeko bi balio erabiltzea erabaki da, horrela atalase mota bakoitzeko atalase gogor bat eta atalase malgu bat probatzeko:

- **Atalase lokalak:**
  - *TOP1 (Atalase lokal gogorra):* Sarrera bakoitzeko antzekotasun balio onena duen itzulpena bakarrik hartzen da ontzat.
  - *TOP3 (Atalase lokal malgua):* Sarrera bakoitzeko 3 antzekotasun balio onenak dituzten 3 itzulpenak hartzen dira ontzat.
- **Atalase globalak:**
  - *0,1 (Atalase global gogorra):* Balio hori baino antzekotasun balio altuagoa duten itzulpen bikoteak bakarrik hartzen dira ontzat.
  - *0,05 (Atalase global malgua):* Balio hori baino antzekotasun balio altuagoa duten itzulpen bikoteak bakarrik hartzen dira ontzat.

Bestalde, eta *IC* teknikarekin buruturiko esperimentuetan egin bezala, *DS* teknikak itzultzen dituen sarrerei sarrera monosemikoak ere gehitu zaizkie esperimentua burutzerakoan, izan ere, gogoratu sarrera monosemiko horiek beti zuzentzat hartzen direla lan honetan. Erabaki hori hartzearen arrazoi nagusia ahalik eta lagin handienarekin lan egin ahal izatea da.

Sarrera bakoitzaren erabilpen maiztasunak *DS* teknikaren portaeran eta lortu den hiztegi elebidunaren izaeran duen eragina aztertu ahal izateko, doitasun eta estaldura balio guztiak sarreren erabilpen maiztasunaren arabera neurtu dira. 5.8 irudian goian aipatu diren atalaseak aplikatuz *DS* teknika gainditzen duten sarreren bataz besteko doitasuna ikus daiteke, maiztasunaren arabera.

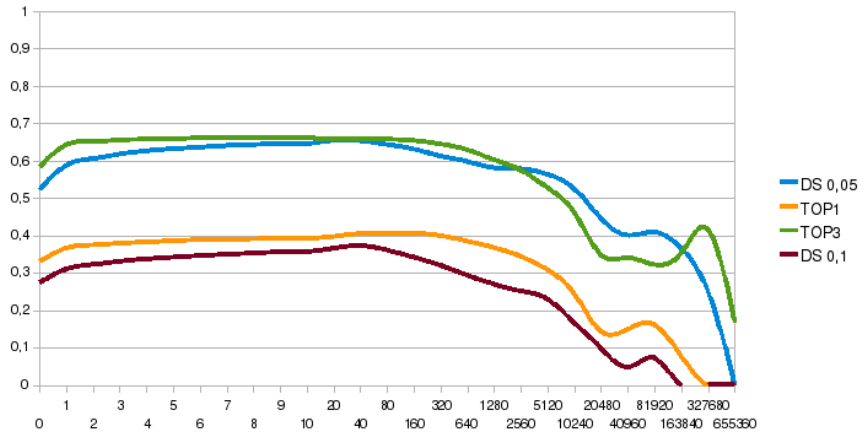


Irudia 5.8: *DS* teknikaren atalase ezberdinak erabiliz lorturiko sarreren bataz besteko doitasuna maiztasunarekiko

5.8 irudian ikus daitekeen moduan, doitasunaren ikuspuntutik ez dago diferentzia handirik atalase lokalen eta atalase globalen bidez diren emaitzen artean. Atalase lokalen nahiz globalen kasuan, baldintza gogorragoak finkatzen direnean (*TOP1* eta *0,1*), doitasunean igoera txiki bat gertatzen dela ikus daiteke (0,4-tik 0,5-era). Bestalde, emaitzak sarreren maiztasunaren arabera aztertzen badira, edozein atalase erabiliz maiztasunak gora egin ahala emaitza doitasun hobea lortzen dela ikus daiteke, maiztasun balio jakin bat arte (f<1.000). Puntu horretatik aurrera, doitasunak beherakada nabarmen bat jasaten du bai atalase globalak aplikatzean eta baita atalase lokal gogorra erabiltzean ere. Datuen arabera, maiztasun altuenetan (f>1.000) portaera hoberena duena atalase lokal malgua da *(TOP3)*, maiztasun txikienetan erakusten duen igoera joera erakusten jarraitzen duelarik maiztasun altuenetan ere. Hortaz, orokorrean maiztasun ezberdinetan portaera konstanteena erakusten duen atalasea *TOP3*-a dela esan daiteke. Bukatzeko, argitu beharra dago mugako maiztasunetan (f>81.920) ageri diren igoera eta jaitsiera handiek ez dutela informazio esanguratsurik eskaintzen, izan ere, maiztasun muga hori gainditzen duten sarreren kopurua oso baxua da (20 sarrera baino gutxiago).

DS teknikaren bidez lortzen diren hiztegietako sarreren bataz besteko estaldurari dagokionez, lortutako emaitzak 5.9 irudian agertzen dira.

HAP masterra

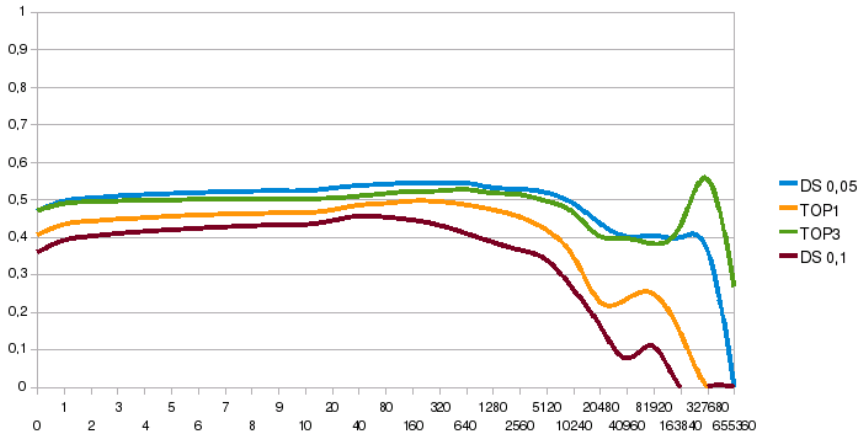


Irudia 5.9: *DS* teknikaren atalase ezberdinak erabiliz lorturiko sarreren bataz besteko estaldura maiztasunarekiko

5.9 irudian ageri den moduan, doitasunaren kasuan gertatzen den bezala sarreren bataz besteko estalduraren kasuan ere ez dago diferentzia handirik atalase lokal eta globalen artean. Atalasearen gogortasuna aztertzen bada, atalase malguenak erabiliz (*TOP3* eta *0,005*) estaldura hobea lortzen dela nabarmena da (0,3-tik ia 0,6-rakoa). Emaitzak sarreren maiztasunaren arabera aztertuz gero, estaldura maiztasun txikienetan nahiko konstante mantentzen bada ere f>50 den kasuetan estalduraren balioak beheraka bat jasaten du. Kasu horietan, atalase global malgua da *(0,05)* emaitza onargarrienak ematen dituena (beherakada geratzen da baina modu kontrolatuagoan). Hori gertatzearen arrazoia maiztasun handiena duten hitzen polisemia mailan bilatu behar da. Jakina da maiztasunak gora egin ahala, sarreren polisemia mailak ere gora egiten duela. Polisemia mailak gora egin ahala (edo itzulpen gehiago izateko aukerak gora egin ahala), zailagoa bihurtzen da beraien testuinguruak ondo adieraztea, eta horregatik atalase global gogorrak *(0,1)* malguak *(0,05)* baino emaitza okerragoak itzultzen ditu. Bestalde, atalase lokalak finkatzen duen ranking bidezko itzulpen aukeraketa mugatua (atalase malguarekin 3 itzulpen eta gogorrarekin bakarra) baldintza gogorregia bihurtzen da aipaturiko polisemia maila altuaren ondorioz (itzulpen zuzen asko egon arren *DS* balio ziurrena dutenak bakarrik hartzen direlako). Arazo honen soluzio posible bat maiztasunak gora egin ahala atalase lokalen rankingak onartzen duen itzulpen kopurua handitzea litzateke. Bukatzeko, eta doitasunaren kasuan gertatu bezala, argitu beharra dago mugako maiztasunetan (f>81.920) ageri diren igoera eta jaitsierek ez dutela informazio esanguratsurik eskaintzen, izan ere, maiztasun muga hori gainditzen duten sarreren kopurua oso baxua da (20 sarrera baino gutxiago).

Doitasuna eta estalduraren arteko erlazioa definitzen duen F-score neurriaren emaitzak 5.10 irudian ikus daitezke, berriro ere maiztasunaren arabera banatuak.

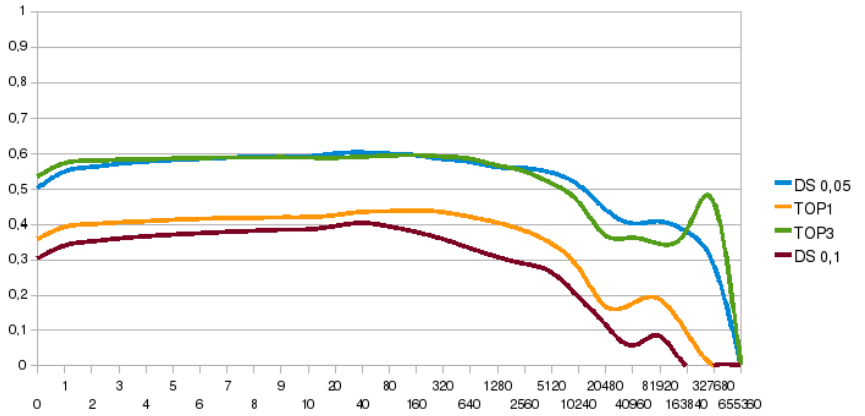
HAP masterra



Irudia 5.10: *DS* teknikaren atalase ezberdinak erabiliz lorturiko sarreren bataz besteko F-scorea maiztasunarekiko

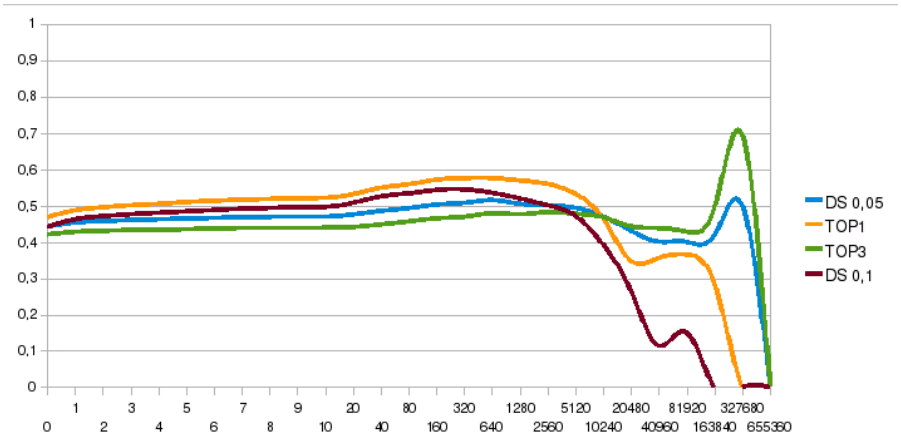
F-scorearen kasuan ere ez dago diferentzia handirik atalase lokal eta globalen artean. Izan ere, aurreko puntuetan ikusi bezala atalase mota bakoitzak portaera hobea du doitasunean edo estalduran (atalase lokalak doitasun hobea eta atalase globalak estaldura hobea). Bestalde, F-scoreak maiztasunarekiko duen aldakortasuna doitasunean eta estalduran ikus daitekeena baino txikiagoa da, bi neurrien konbinaketak aldakortasun hori minimizatzen baitu. Doitasunaren eta estalduraren kasuan gertatu bezala, maiztasun altuenetan lortzen diren emaitzen kalitateak zertxobait behera egiten du.

Bestalde, 4. atalean azaldu bezala, interesgarria da *DS* teknikaren bidez sortzen diren hiztegien emaitzak testuinguru ezberdinetan izan dezaketen erabilgarritasuna estimatzea. Horretarako, F-score neurriaren bi aldaera neurtu dira: estaldurari lehentasuna ematen dion bat (F_2) eta doitasunari garrantzia ematen dion beste bat ($F_{0,5}$). Bi neurri horien emaitzak 5.11 eta 5.12 irudietan ikus daitezke.



Irudia 5.11: *DS* teknikaren atalase ezberdinak erabiliz lorturiko sarreren bataz besteko F_2 maiztasunarekiko

HAP masterra

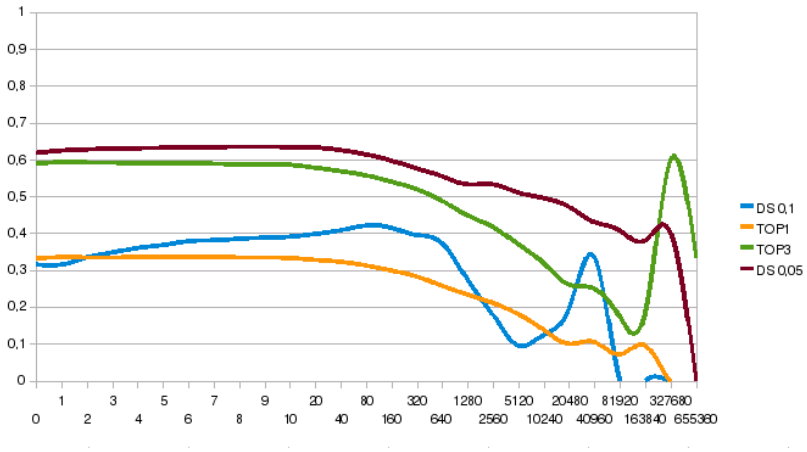


Irudia 5.12: *DS* teknikaren atalase ezberdinak erabiliz lorturiko sarreren bataz besteko $F_{0,5}$ maiztasunarekiko

F-score neurriaren kasuan gertatzen den bezala, atalase moten eta horien balioen araberako emaitzen maiztasunarekiko aldakortasuna oso txikia da. F_2 neurriaren kasuan, estaldurari doitasunari baino pisu handiagoa ematen zaionez atalase malguek emaitza hobeak lortzen dituzte (0,5-tik gora) atalase gogorrek baino, atalase malguen bidez lortzen den estaldura beti altuagoa baita. Aldiz, $F_{0,5}$ neurriaren kasuan, emaitzak oso parekoak dira atalase guztiekin. Hori doitasun balioetan atalase ezberdinen artean dauden diferentziak estalduran daudenak baino txikiagoak direlako gertatzen da, doitasun eta estalduraren arteko oreka bat sortzen delarik doitasunari garrantzi handiagoa ematen zaionean.

DS teknikaren bidez lortutako sarreretan ebaluatu daitekeen beste alderdi bat ematen dituen itzulpenen izaera da, hau da, ematen dituen itzulpenak komunenak diren (ohikoenak) edota aldiz arraroagoak diren (ez hain erabiliak edo ez ohikoak). 4.2.6.2.3 puntuan azaldu bezala, ebaluazio hori burutu ahal izateko gurutzaketa-prozesuaren ostean sorturiko hiztegi elebiduneko itzulpen bikoteek corpus paralelo batean duten agerpenean oinarritzen gara, corpusean gehien agertzen diren itzulpenak komunenak direla pentsa baitaiteke. Ebaluaziorako erabiliko den erreferentziaren abiapuntua gurutzaketatik sorturiko hiztegi zaratatsua denez, kasu honetan ebaluatu daitekeen faktore bakarra estaldura da (sarrera bakoitzeko itzulpen probableenetatik zenbat ematen diren), izan ere, gurutzaketaren bidez lorturiko hiztegiko bikote guztiak laginean sartzeak doitasunaren kalkulua zentzugabe bihurtzen du (itzulpen bikote guztiak ontzat emango liratekeelako).

HAP masterra



Irudia 5.13: *DS* teknikaren atalase ezberdinak aplikatuz lorturiko sarreren itzulpen probableenen estaldura maiztasunarekiko

5.13 irudian ageri den bezala, itzulpen probableenen estaldurak gainerako neurriek erakusten duten antzeko joera erakusten du: maiztasun jakin bat arte nahiko konstante mantentzen da (f<100) baina gero beherakada bat gertatzen da. Atalase globalen eta lokalen artean ez dago diferentzia handirik. Aldiz, emaitzak atalasearen gogortasunaren arabera aztertzen badira, atalase malguen bidez estaldura hobea lortzen dela nabarmena da (0,3-tik 0,6rako igoera). Kasu guztietan aipagarria da sarreren maiztasunak gora egin ahala itzulpen probableenen estaldurak beherakada nabarmena jasaten duela. Honen arrazoia maiztasun handiena duten sarreren polisemia maila altua da, izan ere, sarrera horiek gero eta itzulpen gehiago izateak gero eta gehiago zailtzen du sarrera horien testuinguruak behar bezala adieraztea. Hori dela eta, 5.13 irudian nabaria da atalase malguenek portaera hobea erakusten dutela maiztasun altuenetan ere. Dena den, esan daiteke *DS* teknikak kasu askotan ez dituela itzulpen probableenak itzultzen.

Beraz, egindako esperimentuan oinarrituz, esan daiteke *DS* teknikaren ezaugarri garrantzitsuena testuinguru ezberdinetara egokitzeko duen gaitasuna dela. Izan ere, itzulpen bikoteak balidatzeko aplikatzen diren atalase balio ezberdinen bidez, emaitza onargarriak lor daitezke bai doitasuna garrantzitsua den testuinguruetan eta baita estaldura garrantzitsuagoa den testuinguruetan ere. Bestalde, tratatzen diren sarreren maiztasunak teknikaren emaitzetan duen eragina aztertzeko eginiko esperimentuei esker, nabaria da *DS* teknika nahiko sentikorra dela sarreren maiztasunarekiko. Horrela, maiztasun txiki eta ertainetako sarreretan teknikaren portaera konstantea izan arren, maiztasun altueneko sarreren kasuan emaitzek okerrera egiten dute. Fenomeno hori maiztasun altuenetako sarrerek izan ohi duten polisemia maila altuaren ondorioz gertatzen da, izan ere itzulpen asko izatearen ondorioz horien testuinguruak behar bezala adieraztea zailagoa bihurtzen da, itzulpen guztien testuinguruak haien artean difuminatzen baitira. Maiztasunaren eragin hori itzulpen probableenen tratamenduan ere nabari da, emaitzetan ikusi bezala maiztasun altuenetako sarreren kasuan itzulpen probableenen estaldurak nabarmen behera egiten baitu. Beraz, *DS* teknika testuinguru ezberdinetara moldatzeko egokia da, baina maiztasunarekiko sen-

tikortasun nabarmena du.

5.1.4 *IC* eta *DS* tekniken konbinaketa

Aurreko kapituluko 4.2.5 puntuan azaldu bezala, lan honetan *IC* eta *DS* tekniken emaitzak konbinatzea proposatzen da, paradigma eta baliabide ezberdinetan oinarritzen diren teknikak izanik bien emaitzek dibergentzia maila minimo bat izateko aukera baitago. Horretarako, bi tekniken emaitzak bi eratara konbinatzeko aukera aurkezten da: konbinaketa gordina erabiliz *(UNION)* edota konbinaketa lineala aplikatuz *(LCOMB)*.

Orain arte *IC* eta *DS* tekniken bidez lortu diren emaitzekiko konparaketa burutu ahal izateko, orain arte aurkeztu diren testuinguru ezberdinetan bi tekniken konbinaketak duen portaera orokorra neurtuko da. Era honetara, *IC* eta *DS* teknikak modu isolatuan erabiliz lortzen diren emaitzak gainditzeko aukera aztertu nahi da. Konparaketa burutu ahal izateko aukeraturiko testuinguruak simulatzen dituzten neurriak *IC* eta *DS* tekniken emaitzak aztertzeko erabili diren F, F_2 eta $F_{0,5}$ dira. Bestalde, emaitzak sarreren maiztasun-faktorearen arabera aztertu ahal izateko wF *(weighted F)* neurria erabiltzea proposatzen da. Neurri honek sarrera bakoitzari pisu jakin bat esleitzen dio erabilpen maiztasunaren arabera. Horrela, gero eta sarrera maiztasun handiagoa izan orduan eta pisu handiagoa esleitzen zaie, maizen erabiltzen diren sarrerak hobeto adieraztea propietate desiragarria baita edozein hiztegiren kasuan. 5.3 taulan konbinaketa mota ezberdinen bidez lorturiko emaitzak ageri dira *DS* eta *IC* teknikekin lorturiko emaitzekin batera. Konbinaketaren kasuan, neurri bakoitzarekin lorturiko emaitzak optimizatu egin dira, hau da, aplikatzen diren atalase balioak emaitza maximoak lortzen dituztela bermatu da. Atalase horien balio optimoak estimatzeko 5.1.3 puntuan azalduriko metodologia jarraitu da, hau da, *DS* teknikaren kasuan aplikatzen diren atalase balioak *IC* teknikak itzultzen dituen emaitzen bidez doitu dira.

Metodoa	F	wF	F_2	$F_{0,5}$
IC	0,34	0,27	0,27	0,46
DS	0,47	0,44	0,64	0,46
UNION	0,52	0,49	0,65	0,49
LCOMB	0,52	0,49	0,67	0,52

Taula 5.3: *IC* eta *DS* tekniken konbinaketaren bidez lorturiko emaitzen konparaketa jatorrizko balioekiko

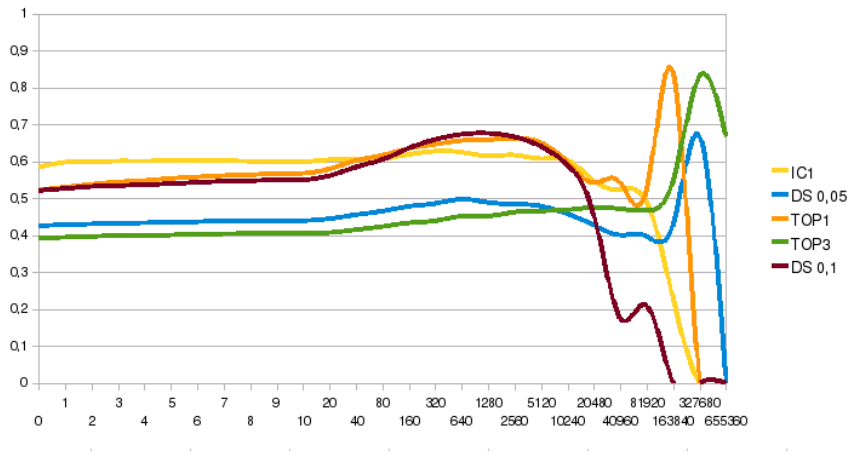
5.3 taulan ikus daitekeen bezala, *IC* eta *DS* tekniken konbinazioaren bidez bi teknika horiek isolaturik aplikatzean lortzen diren emaitzak gainditzea lortzen da. Hobekuntza hori gainera erabilpen testuinguru ezberdinak adierazteko erabili diren neurri guztietan nabaritu daiteke. *IC* teknikarekin lorturiko emaitzekiko, neurri gehienetan hobekuntza nabarmena lortzen da. Emaitza antzekoenak doitasunari lehentasuna ematen zaion testuinguruan lortzen dira ($F_{0,5}$ neurrian 0,03 hobekuntza), gainerako guztietan hobekuntza oso nabarmena delarik. *DS* teknikaren kasuan lortzen den hobekuntza ez da *IC* teknikarekiko lortzen dena bezain handia baina hala ere hobekuntza minimo bat nabari da neurri

guztietan (0,01-etik 0,05 arte neurriaren arabera). Konbinaketa egitekorakoan aplikaturiko bi metodologiak konparatuz, emaitzetan alde txikia dagoela erakusten da. Dena den, doitasuna edo estaldurari lehentasuna ematen zaion testuinguruetan konbinaketa linealak *(LCOMB)* emaitza hobeak lortzen dituela ikus daiteke. Hobekuntza honen zergatia konbinaketa linealak teknika bakoitzari esleitzen dion pisuan dago. Izan ere, konbinaketan erabiltzen den teknika bakoitzak portaera hobea erakusten du neurri batean bestean baino (*IC*-k doitasun hobea lortzen du eta *DS*-k aldiz estaldura hobea). Konbinaketa linealak aukera ematen du erabilpen testuinguru bakoitzean ($F_{0,5}$ edo F_2 neurrietan alegia) *IC* teknikari edo *DS* teknikari pisu handiagoa esleitzeko, eta ondorioz neurri bakoitzean teknika bakoitzaren emaitzak modu optimoan konbinatzea ahalbidetzen du. Beraz, esan daiteke konbinaketa lineala *(LCOMB)* testuinguru ezberdinetara konbinaketa gordina *(UNION)* baino hobeto egokitzen dela.

5.1.5 Tekniken arteko konparaketa

Esperimentuak burutzeko erabilitako teknika bakoitzaren emaitzak aztertu ostean, jarraian emaitza horien konparaketa bat egingo da teknika bakoitzak egin dezakeen ekarpena neurtzeko. Horretarako, orain arte emaitzak aztertzeko aurkeztu den neurri bakoitzeko konparaketa egingo da.

Teknika bakoitzaren bidez lorturiko sarreren bataz besteko doitasunari dagozkion emaitzak 5.14 irudian ageri dira, maiztasunarekiko adierazita.

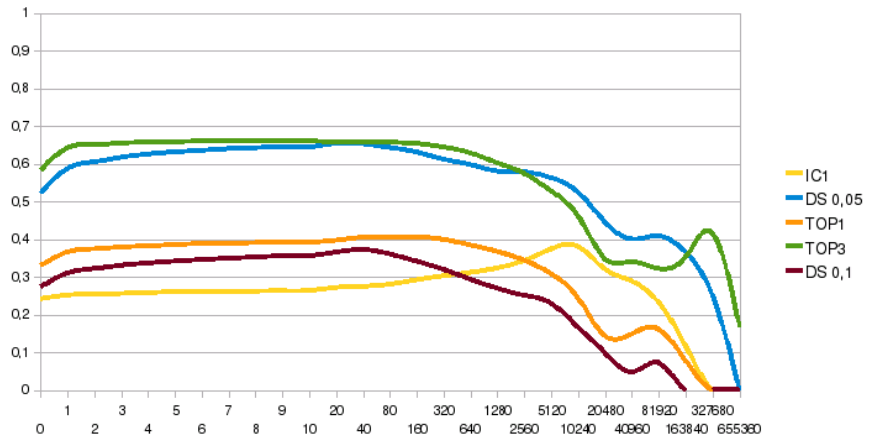


Irudia 5.14: *DS* eta *IC* tekniken bidez lorturiko sarreren bataz besteko doitasun konparaketa maiztasunarekiko

Lorturiko emaitzen arabera, ez dago diferentzia handirik *IC* teknikaren eta atalase gogorrak aplikatuz *DS* teknikaren bidez lorturiko sarreren bataz besteko doitasunean. Gainera, kasu guztietan maiztasunaren arabera doitasunak erakusten duen joera mantendu egiten da (hasierako igoera eta maiztasun altuenetako beherakada), nahiz eta *IC* teknikaren maiztasunarekiko aldakortasuna *DS* teknikarena baina txikiagoa dela esan daitekeen

(konstanteago mantentzen da maiztasun aldaketarekiko). Beraz, bataz besteko doitasunari dagokionez *IC* eta *DS* teknikaren bidez lortzen diren emaitzak parekoak dira.

Teknika bakoitzaren bidez lorturiko sarreren bataz besteko estaldurari dagozkion emaitzak 5.15 irudian ikus daitezke, sarreren maiztasunarekiko erakutsiak.

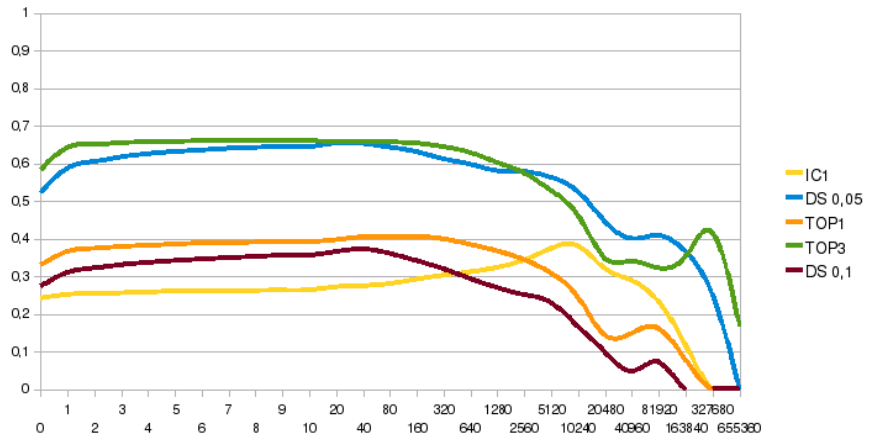


Irudia 5.15: *DS* eta *IC* tekniken bidez lorturiko sarreren bataz besteko estaldura konparaketa maiztasunarekiko

5.15 irudian ikus daitekeen bezala, *DS* teknikaren bidez lorturiko bataz besteko estaldura *IC* teknikaren lor daitekeena baino altuagoa da (ia 0,4-ko diferentzia), atalase malguak aplikatzen diren kasuetan. Izatez, *IC* teknikaren estaldura *DS* teknikarena baino zertxobait baxuagoa da atalase gogorrenak erabiltzen diren kasuetan ere. *IC* teknikak estalduraren kasuan lortzen dituen emaitza kaxkarrak teknikak finkatzen dituen baldintzen gogortasuna erakusten dute (emaitzen arabera baldintza horiek *DS* teknikan atalase gogorrenak aplikatzea baino zorrotzagoak dira). Bestalde, emaitza horiek sarreren maiztasunaren arabera aztertzen badira, nabaria da *IC* teknikak portaera hobea erakusten duela maiztasun altuagoetan (f>50) *DS* teknikak baino. Aurreko puntuetan azaldu den bezala, maiztasun altuena duten testuinguruak behar bezala adieraztea zaila bihurtzen da eta hori dela eta *DS* teknikaren bidez lorturiko emaitzak okerrera egiten dute maiztasunak gora egin ahala. Aldiz, *IC* teknika gurutzaturiko hiztegien egituran bakarrik oinarritzen denez, faktore honek ez dio eragiten, are gehiago, gero eta itzulpen gehiago izateak (polisemia maila altuagoak) sendotasun handiagoa ematen dio *IC* teknikari, izan ere, gogoratu teknikaren funtsa sarrera eta itzulpenen arteko hiztegi loturak direla. Beraz, gero eta itzulpen gehiago izan, orduan probabilitate altuagoa dago hiztegi lotura gehiago eta sendoagoak lortzeko. Argitu beharra dago maiztasun altuenetan ageri diren beherakadak ez direla esanguratsuak (f>10.000), maiztasun horietan aztertzen den sarrera kopurua oso baxua baita maiztasun baxuagoetan aztertzen direnekiko. Beraz, estaldurari dagokionez oro har *DS* teknikaren bidez emaitza askoz hobeak lor daitezke. Hala ere, maiztasun altueneko sarreren kasuan *IC* teknikaren portaera *DS* teknikarena baino hobea dela nabaria da.

Bi tekniken doitasuna eta estalduraren arteko erlazioa definitzen duen F-score neurria-

ren emaitzak 5.16 irudian ikus daitezke, berriro ere maiztasunaren arabera banatuak.

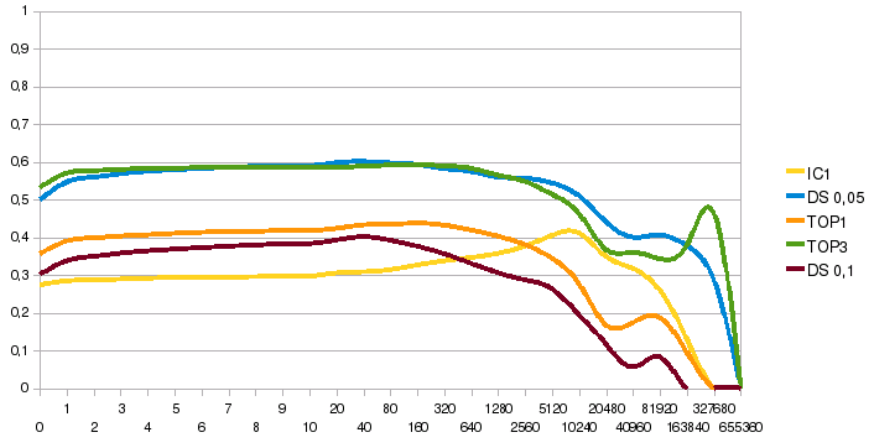


Irudia 5.16: *DS* eta *IC* tekniken bidez lorturiko sarreren bataz besteko F-score konparaketa maiztasunarekiko

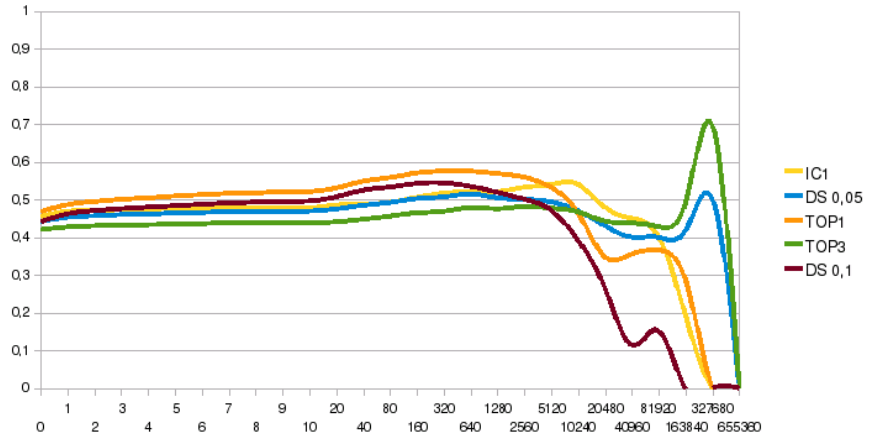
5.16 irudian ikus daitekeen bezala, *DS* teknikak *IC* teknikak baino F-score hobea lortzen du atalase malguak nahiz gogorrak erabiliz. *DS* teknikaren kasuan, emaitza horiek zertxobait hobeak dira atalase malguak aplikatzen direnean, diferentzia honen eragile nagusia atalase malguenen bidez lortzen den bataz besteko estaldura altua delarik. Emaitzak sarreren maiztasunaren arabera aztertzen badira, maiztasun baxuenetan bi teknikek antzeko joera jarraitzen dutela ikus daiteke (f<100). Maiztasun altuagoetan aldiz, *IC* teknikaren emaitzek hobera egiten dute maiztasunak gora egin ahala, *DS* teknikarenak beherakada nabarmena jasaten duten bitartean. Aurreko puntuan azaldu bezala, diferentzia hori teknika bakoitzaren bataz besteko estaldurak eragiten du, izan ere *IC* teknikak maiztasun altueneko sarrerak *DS* teknikak baino hobeto tratatzeko joera erakusten du. Beraz, F-scoreari dagokionez oro har *DS* teknikaren bidez emaitza hobeak lor daitezke. Argitu beharra dago maiztasun altuenetan ageri diren beherakadak ez direla esanguratsuak (f>10.000), maiztasun horietan aztertzen den sarrera kopurua oso baxua baita maiztasun baxuagoetan aztertzen direnekiko. Edozein kasutan, maiztasun altueneko sarreren kasuan *IC* teknikaren portaera *DS* teknikarena baino hobea dela nabaria da.

Bestalde, 4. atalean azaldu bezala, interesgarria da bi tekniken bidez sortzen diren hiztegien emaitzak testuinguru ezberdinetan izan dezaketen erabilgarritasuna konparatzea. Horretarako, F-score neurriaren bi aldaera neurtu dira: estaldurari lehentasuna ematen dion bat (F_2) eta doitasunari garrantzia ematen dion beste bat ($F_{0,5}$). Bi neurri horien emaitzak 5.17 eta 5.18 irudietan ikus daitezke.

HAP masterra



Irudia 5.17: *DS* eta *IC* tekniken bidez lorturiko sarreren bataz besteko F_2 konparaketa maiztasunarekiko

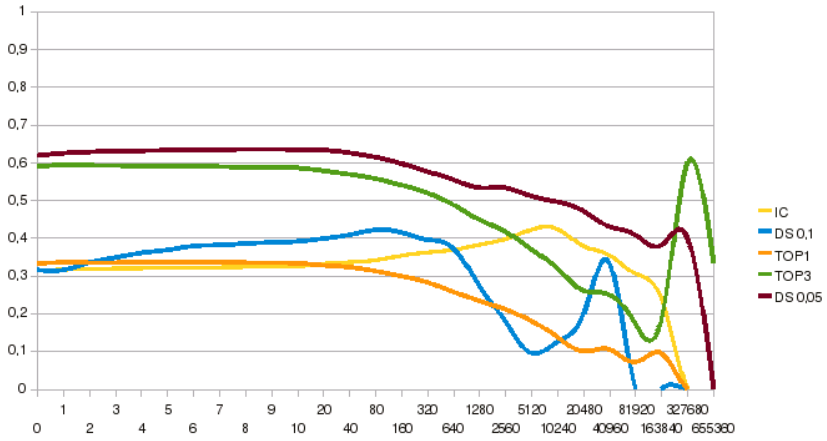


Irudia 5.18: *DS* eta *IC* tekniken bidez lorturiko sarreren bataz besteko $F_{0,5}$ konparaketa maiztasunarekiko

5.17 irudian ageri den bezala, estaldurari doitasunari baino garrantzia handiagoa ematen zaion kasuan (F_2), *DS* teknikaren bidez lortzen diren emaitzak *IC* teknikaren bidez lortzen direnak baino hobeak dira. Hobekuntza hori nabarmenagoa da atalase malguak aplikatzen diren kasuetan (0,25-ko hobekuntza). Bestalde, doitasunari pisu handiagoa esleitzen zaion testuinguruetan ($F_{0,5}$), bi tekniken bidez lortzen diren emaitzak oso antzekoak dira, aurreko puntuetan ikusi bezala, bi tekniken bidez antzeko doitasun maila lortzen baita. Kasu guztietan, aipagarria da maiztasun altuenetan bi teknikek duten portaera ezberdina: *IC* teknikaren emaitzek hobera egiten dute sarreren maiztasunak gora egin ahala eta aldiz *DS* teknikaren emaitzek okerrera egiten dute. Beraz, esan daiteke orokorrean *DS* teknikak eszenatoki ezberdinetara egokitzeko duen ahalmena handiagoa dela *IC* teknikak

duena baino, baina maiztasun altuenetan aldiz *IC* teknikak *DS* teknikak baino portaera hobea erakusten du.

Bi tekniken bidez lorturiko emaitzen konparaketarekin amaitzeko aztertu den azken faktorea sarrera bakoitzeko ematen diren itzulpenen izaera da, hau da, ematen dituzten itzulpenak komunenak diren (ohikoenak) edota aldiz arraroagoak diren (ez hain erabiliak edo ez ohikoak). Teknika bakoitzaren kasuan buruturiko analisiaren bidez lorturiko emaitzak 5.19 irudian ageri dira konbinaturik.



Irudia 5.19: *DS* eta *IC* tekniken bidez lorturiko sarreren itzulpen probableenen estaldura konparaketa maiztasunarekiko

Kasu honetan ere orokorrean *DS* teknikaren bidez lorturiko emaitzek *IC* teknikaren bidez lortutakoak gainditzen dituzte. Bi tekniken bidez lorturiko emaitzak antzekoak dira *DS* teknikan atalase gogorrenak aplikatzen direnean, 5.1.3 puntuan azaldu bezala atalase horiek portaera okerragoa erakusten baitute itzulpen probableenak itzultzerakoan. Atalase malguak aplikatzen diren kasuetan, *DS* teknikaren bidez lortzen diren emaitzak 0,3-an hobetzen dituzte *IC* teknikaren bidez lortutakoak. Kasu honetan ere, aipagarria da maiztasunarekiko emaitzek duten aldakortasuna: *IC* teknikaren emaitzek hobera egiten dute sarreren maiztasunak gora egin ahala eta aldiz, *DS* teknikaren bidez lorturiko emaitzek okerrera egiten dute. Lehenago aipatu bezala, aldakortasun honen eragile nagusia maiztasun altueneko sarreren polisemia maila da: itzulpen asko izateak zaildu egiten baitu *DS* teknikan beharrezkoak diren testuinguruen kalitatea bermatzea baina aldiz onuragarria da *IC* teknikak oinarritzat hartzen dituen hiztegi loturen kopurua eta sendotasuna handitzeko. Beraz, eta azterturiko gainerako neurrietan nabaritu bezala, orokorrean *DS* teknikak itzulpen probableenen tratamendu hobea egiten du *IC* teknikak baino, baina maiztasun altuenetan *IC* teknikaren bidez lorturiko emaitzak fidagarriagoak dira.

5.1.6 Esperimentuaren ondorioak

Euskara-Gaztelera hiztegi elebidunarekin egindako esperimentuetatik hainbat ondorio atera daitezke. Erabili eta aplikatu diren teknikei dagokionez, portaera oso ezberdina dutela

nabaria da. *IC* teknikaren kasuan, garbi geratu da batez ere doitasunak garrantzia duen egoeratara egokitzen dela, lortzen dituen estaldura-emaitzak benetan kaxkarrak baitira. Teknikak emaitza erabilgarriak eskain ditzan, prozesuan parte hartzen duten hiztegiek sarrera bakoitzeko ahalik eta itzulpen gehien ematea exijitzen du. Anbiguetate maila altu hori mantentzea oso baldintza gogorra bihurtzen da sarrera askoren kasuan eta ondorioz emaitzen estaldura nahiko kaxkarra izan ohi da. Aldiz, *DS* teknika askoz hobeto egokitzen da edozein testuingurura, aplikatu ahal diren atalase ezberdinei esker. Horrela, doitasunari dagokionez, *IC* teknikak lortzen dituen emaitzak parekatzea lortzen da, eta estaldurari dagokionez *IC* teknikaren bidez lorturiko emaitzak izugarri hobetzen ditu. *DS* teknikaren ahulezia bakarra sarreren maiztasunarekiko duen sentikortasuna da, maiztasun altueneko sarreren kasuan orokorrean emaitzek okerrera egiten dutelarik (*IC* teknikan ez bezala). Hala ere, kontuan eduki behar da *DS* teknika *IC* teknika baino baliabide gehiagotan oinarritzen dela, corpus konparagarri baten erabilera beharrezkoa baita. Hortaz, baliabideen ikuspunturik *IC* teknikak baino baldintza gogorragoak exijitzen ditu.

Bestalde, esperimentu honen beste helburuetako bat *DS* eta *IC* tekniken emaitzak konbinatzeko aukera aztertzea izan da. 4.2.5 puntuan azaldu bezala, bi teknika horiek paradigma eta baliabide ezberdinetan oinarritzen direnez, lortzen dituzten itzulpen bikoteen multzoetan zenbateko dibergentzia maila dagoen aztertu da. Horrela, *IC* eta *DS* teknikak konbinatuz lorturiko emaitzek teknika bakoitzaren bidez lorturikoak gainditzen dituztela ikusi da, eta beraz, bi tekniken bidez lorturiko emaitzetan dibergentzia maila minimo bat dagoela frogatu da. Tekniken konbinaketaren bidez lorturiko emaitzek gainerako emaitza guztiak gainditzen dituzte. Dena den, kontuan hartu behar da konbinaketa hori aplikatu ahal izateko bai *IC* eta bai *DS* teknikak exijitzen dituzten baliabide behar guztiak bete beharra dagoela.

Azkenik, esperimentuetan sorturiko hiztegi baliabidearen erabilgarritasuna aztertu beharra dago. Lorturiko emaitzetan oinarrituz, nabaria da hainbat testuinguruetan erabili ahal izateko zuzenketa lan handia behar duela (kontsulta hiztegi bezala erabiltzeko adibidez). Hala ere, mota honetako baliabideen sorkuntzan laguntza bezala erabil daiteke, adibidez sarrera edo itzulpen berriak proposatzeko. Gainera, hizkuntza naturalen prozesamenduaren arloko hainbat atazatan ere erabilgarria izan daiteke (itzulpen automatikoan adibidez). Esperimentu horien hurrengo urratsetako bat sorturiko hiztegiak testuinguru horietan duen erabilgarritasuna frogatzea izango litzateke, proba eta testak burutuz.

5.2 Euskara-Txinera hiztegia

Puntu honetan Euskara-Txinera hiztegi elebidunaren sorkuntza automatikoak emandako emaitza guztiak azalduko dira. Horretarako, esperimentuaren fase bakoitzeko emaitzak puntu ezberdinetan banatuko dira: lehendabizi gurutzaketatik sorturiko hiztegi gordinaren datuak emango dira eta jarraian teknika bakoitzaren portaera eta horien bidez lorturiko emaitzak azalduko dira (*IC*, *DS* eta *IC-DS* konbinaketa). Bukatzeko, azken puntuetan tekniken konparaketa bat eta esperimentuaren ondorio orokor batzuk azalduko dira.

5.2.1 Gurutzaketa-prozesua

Euskara-Txinera hiztegi elebiduna automatikoki sortzeko prozesuan 3.2 atalean definituriko *EU-EN* eta *EN-ZH* hiztegiak erabili dira. Horretarako, lehendabizi hiztegi horien edukiak normalizazio-prozesutik pasatu dira bi hiztegien edukien formatuak ahalik eta antzekoenak bihurtzeko. 4.2.1 atalean azaldu bezala, prozesu honen helburua gurutzaketaren emaitzak maximizatzea da. *EU-EN* eta *EN-ZH* hiztegi normalizatuen ezaugarri orokorrak eta gurutzaketaren bidez lorturiko *EU-ZH* hiztegi berriaren propietate orokorrak 5.4 taulan ikus daitezke.

Hiztegia	Sarrera kopurua	Itzulpen kopurua	Anbiguotasun maila
EU-EN (jatorrizkoa)	17.699	42.994	2,43
EN-ZH (jatorrizkoa)	37.313	63.899	1,71
EU-ZH (gurutzaketa)	13.544	182.089	13,44

Taula 5.4: *EU-ZH* hiztegiaren gurutzaketa-prozesuko hiztegien propietate orokorrak

5.4 taulako datuetan ikus daitekeen bezala, *EN-ZH* hiztegiaren sarrera eta itzulpen kopurua *EU-EN* hiztegiarena baino askoz ere altuagoa da (gutxi gora-behera 20.000 sarrera eta itzulpen gehiago). Honen ondorioz, gurutzaketaren ostean sortzen den hiztegiaren tamaina *EU-EN* hiztegiaren tamainak mugatzen du. Horrela, gurutzaketaren ostean sortzen den hiztegiaren sarrera kopurua jatorrizko hiztegiena baino txikiagoa da (*EU-EN* hiztegiak baino 4.000 sarrera gutxiago). Hala ere, lortzen den itzulpen kopurua izugarria da proportzioan, jatorrizko hiztegien itzulpen kopurua ia hirukoiztera heltzen delarik. Itzulpen kopurua hain altua izatearen ondorioz, gurutzaketaren bidez sorturiko hiztegiko sarreren bataz besteko itzulpen kopurua 13,44-koa da, hau da, sarrera bakoitzeko bataz beste 13,44 itzulpen ematen dira. 5.4 taulan ikus daitekeen bezala, kopuru hori jatorrizko hiztegietan baino askoz altuagoa da. Hiztegiaren analisi sakonago bat eginez, sarreren %85,58-ak (11.591 sarrera, 180.136 bikote) itzulpen posible bat baino gehiago du, hau da, polisemikoak dira (gainerakoa %14,42-a itzulpen bakarreko sarrera monosemikoek osatzen dute).

Bestalde, hiztegi honekin burutu diren esperimentuetan sarreren maiztasunak emaitzetan duen eragina aztertu da. Horretarako, 3 maiztasun tarte definitzen dira: maiztasun txikiko tartea ($0<f\leq 20$), maiztasun ertaineko tartea ($20<f<250$) eta maiztasun handiko tartea ($f\geq 250$). 5.5 taulan gurutzaketaren ostean sorturiko hiztegiko sarrera mota ezberdinen kopuruak ikus daitezke maiztasun tartearen arabera taldekatuak.

Sarrerak	$0<f\leq 20$	$20<f<250$	$f\geq 250$	Denak($f>0$)
Monosemikoak	1.065	651	237	1.953
Polisemikoak	4.681	3.530	3.385	11.591
Denak	5.746	4.176	3.622	13.544

Taula 5.5: *EU-ZH* hiztegiaren sarrera kopuruak maiztasun tarteen arabera banatuak

5.2.2 *IC* teknika

Aurreko puntuan sortutako hiztegi gurutzatua eta jatorrizko bi hiztegi elebidunak erabiliz (*EU-ZH* hiztegia eta *EN-ZH* hiztegia) *IC* teknika aplikatu da. 5.6 taulan 4.2.3 puntuan azaldutako irizpideak jarraituz definituriko sarrera motak eta kopuruak ikus daitezke.

Mota	Sar. kop.	Sar. kop. % totalarekiko	Itz. kop.	Itz. kop. % totalarekiko
A_1	1.953	%14,41	1.953	%2,33
A_2	98	%0,72	201	%0,23
B	3.595	%26,54	12.643	%15,05
D_1	1.356	%10,01	2.712	%3,32
D_2	6.542	%48,30	66.457	%79,14
E	4.128	-	-	-

Taula 5.6: Sarrera mota bakoitzeko kopuruak *IC* teknika aplikatu ondoren

Guztira, itzulpen bikoteak dituzten sarrera mota guztietako sarrerak kontuan hartuta (A_1, A_2, B, D_1 eta D_2), 13.544 sarrera eta 83.966 itzulpeneko lagina lortu da. Azpimarratu beharra dago gurutzaketaren ostean lorturiko hiztegiaren itzulpen kopurua (182.089) lagin honetan lortzen dena baina askoz ere handiagoa dela (83.966). Falta diren gainerako itzulpen guztiak (98.123) *IC* teknika aplikatzearen ondorioz galtzen dira, izan ere, eta 4.2.3 puntuan azalduko bezala, sarrera bakoitzeko gutxienez B motako itzulpen bat topatzen denean, sarrera horren D_1 eta D_2 motako itzulpen hautagai guztiak prozesutik kanpo geratzen dira, sailkatzen den elementua sarrera bera baita. Lagin horretatik *IC* teknika gainditzeko baldintzak bete dituzten sarrera bakarrak B motakoak dira. Hortaz, bikote hautagai guztietatik *IC* teknikak finkatzen dituen baldintzak betetzen dituzten sarrerak 3.595 dira (12.643 itzulpen bikote), hau da, gurutzaketa egin ostean lorturiko hiztegiko sarreren %26,54-a bakarrik (bikoteen %7-a). Datu horien arabera, *IC* teknika gainditzeko finkatzen diren baldintzak nahiko gogorrak direla esan daiteke, edo beste era batera esanda, *IC* teknikak bikote hautagai kopuru osoaren zati bat bakarrik tratatu dezake.

IC teknikaren baldintzak gainditzen dituzten sarreren kalitatea neurtu ahal izateko, B motako sarrera eta horien itzulpenak aztertu dira eskuzko ebaluazio-prozesuaren bidez (5.2.4).

5.2.3 *DS* teknika

Aurreko puntuan sortutako hiztegi gurutzatua, jatorrizko bi hiztegi elebidunak eta bi corpus elebakar konparagarriak erabiliz (*EU-EN* hiztegia, *EN-ES* hiztegia, *EU* corpusa eta *ES* corpusa) *DS* teknika aplikatu da. 4.2.4 puntuan azaldu bezala, *DS* teknika aplikatu ahal izateko eman beharreko lehen urratsa testuinguruen itzulpena egiteko erabiliko den hiztegi elebiduna sortu edo aukeratzea da. Aukera guztiekin esperimentatu ostean, anbiguotasunik ez duten sarrerak eta sarrera polisemikoen itzulpen maizenak (helburu hizkuntzako corpusaren arabera) biltzen dituen hiztegia erabiltzea erabaki da, aurreko kapituluan (4.2.4 puntuan) azalduriko aukeren artean emaitza onenak modu honetara lortzen baitira.

DS teknikaren emaitzak aztertu ahal izateko, lehendabizi emaitzak ontzat hartzeko ezarriko den atalasea definitu beharra dago. 4.2.4 puntuan azaldu bezala, lan honetan bi atalase mota erabiliko ditugu, atalase globalak eta atalase lokalak. Bestalde, *EU-ES* hiztegiaren sorkuntza eta esperimentuetan egin bezala, atalase malgu eta atalase gogorren erabilpena aztertu nahi da. Bi kontzeptuak nahastuz, esperimentu honetan bi atalase erabiltzea erabaki da:

- **Atalase malgua *(DSF)*:** Atalase global bat erabiltzen da atalase malgu moduan, zehazki 0,05 atalasea. Horrela, balio hori baino antzekotasun balio altuagoa duten itzulpen bikoteak bakarrik hartzen dira ontzat.
- **Atalase gogorra *(DSS)*:** Atalase lokal bat erabiltzen da atalase gogor moduan, zehazki *TOP3* atalasea. Horrela, sarrera bakoitzeko 3 antzekotasun balio onenak dituzten 3 itzulpenak hartzen dira ontzat. Atalase lokal gogorragoak aplikatzeko aukera egon arren, esperimentu honetarako nahiko gogorra dela estimatu da, batez ere atalase malguarekin konparatuz.

Bi atalase horiek aplikatuz lorturiko sarrera eta horien itzulpenak aztertu dira eskuzko ebaluazio-prozesuaren bidez (5.2.4).

5.2.4 Ebaluazioa

Aurreko kapituluan azaldu bezala, Euskara-Txinera hiztegiaren ebaluazio-prozesua eskuz egitea erabaki da, erreferentzia hiztegi erabilgarri bat ez edukitzeaz gain, azterketa automatikoaren ondorioz sor daitezkeen erroreak ekidin ahal izateko. Horretarako, hiztegi osoaren eskuzko azterketak suposatuko lukeen lan eskerga ekiditeko, emaitza hiztegi honetatik hausaz sortutako lagin bat aztertu da. Lagin honen sorkuntza-prozesua eta emaitzak 5.2.4.1 puntuan azalduko dira. Bestalde, sarreren eta itzulpenen estaldura orokorraren emaitzak ere neurtuko dira puntu berean. Aurreko kapituluko 4.3.6.2.3 puntuan azaldu bezala, neurri horiek ez dute itzulpen bikoteen zuzentasuna kontutan hartzen, gurutzaketa-prozesuan parte hartzen duten sarrera eta itzulpen posibleen multzo osotik emaitza hiztegiak tratatzen duen ehunekoa baizik. Emaitza horiek modu automatikoan neurtzen dira (itzulpena bikoteen zuzentasuna ez baita kontuan hartu behar).

5.2.4.1 Laginaren eskuzko ebaluazio-emaitzak

Euskara-Txinera hiztegi elebiduna sortzeko 4.3.6.2.1 puntuan definituriko irizpideak aplikatu dira. Irizpide horien bidez, *IC* eta *DS* teknikak baldintza berean ebaluatuko direla bermatzeaz gain, sortzen den laginaren propietateak jatorrizko *EU-ZH* hiztegiaren oso antzekoak izango direla ziurtatzen da. Horrela, lagina osatzeko aurretik definituriko maiztasun tarte bakoitzeko ($0<f\leq20$, $20<f<250$ eta $f\geq250$) 50 sarrera aukeratu dira hausaz. Beraz, laginaren hasierako tamaina 150 sarrera eta 3.407 itzulpen bikotekoa da. Lagin honetako itzulpen bikote bakoitzaren azterketa 4.3.6.2.2 puntuan definituriko irizpide eta ebaluazio klaseak erabiliz burutu da (gogoratu itzulpen bikote bakoitzeko bi epai eman

behar direla emaitzaren ziurtasuna bermatzeko). Prozesua 8 lexikografoz osaturiko epaile talde batek burutu du. Epaileen arteko lagin banaketa burutzeko, lagina 8 zati ezberdinetan banatu da, eta zati bakoitza bikoiztu egin da bi epaile ezberdinen epaiak lortu ahal izateko (osotara 16 zati, binaka errepikatuak). Epaile bakoitzari, 16 zati horietatik bi zati ezberdin esleitu zaizkio. 5.7 taulan lagin osoaren gainean egindako epai bikoitzaren adostasun maila erakusten duten emaitzak ikus daitezke, 4.3.6.2.2 puntuan definituriko kategoria-sistemaren arabera banatuak. Taulako errenkadek lehenengo epailearen epaiak adierazten dituzte eta zutabeek bigarren epailearen epaiak.

		B epailea			
		Okerra	Zuzena	Kategoria okerra	Zalantza
A epailea	Okerra	860	612	147	300
	Zuzena		1.184	154	367
	Kategoria okerra			164	44
	Zalantza				75

Taula 5.7: Hausaz sorturiko laginaren azterketaren emaitzen adostasun taula

5.7 taulako emaitzetatik hainbat alderdi aipatu behar dira. Hasteko, epaile ezberdinen arteko desadostasun maila handia azpimarratu behar da. Horrela, batez ere kezkagarrienak diren desadostasun kasuak *zuzena-okerra* (612 itzulpen bikote) eta *zuzena-zalantza* (367 itzulpen bikote) dira, izan ere, desadostasun kasuak izateaz gain ematen diren epaiak oso kontrajarriak daude. Bestalde, epaiak materakoan zalantzazko bezala epaitzen diren itzulpen bikoteen kopuru altua ere kezkagarria da (786 itzulpen bikote osotara). Datu kaxkar horien zergatiak bi izan daitezke. Lehena epaile bezala jokatu duten lexikografoen hizkuntzen ezagutza mugatua da, izan ere, epaile guztiek euskara eta ingeles maila ona izan arren, ez dute txinatar hizkuntzako ezagutzarik. Bestalde, erreferentzia laguntza bezala erabilitako hiztegiak (3.2.5) ematen dituen itzulpenen kalitate mugatua ere arazo horien sortzaile izan daiteke. Esperimentuarekin aurrera jarraitzeko, 4.3.6.2.2 puntuan azaldu bezala adostasun kasu guztiak laginean mantentzen dira eta desadostasun kasu guztiak kanporatu. Desadostasun gogorrenen kasuan *(zuzena-okerra)*, itzulpen bikote horiek berriro aztertzen dira epai komun batera heltzeko helburuarekin. 5.8 taulan, desadostasun gogorrenak errebisatu ondorengo emaitzak eta adostasun kasuekin lorturiko emaitzak batuta lortzen den azken laginaren datuak ikus daitezke. Zalantzazko kasuak ere laginetik kanpo uzten dira.

Epaiak	Okerra	Zuzena	Kategoria okerra
Bikoteak	1.070	1.377	169

Taula 5.8: Desadostasun kasuak garbitu ostean lorturiko laginaren datuak

5.8. taulan ikus daitekeen bezala, esperimentuan zehar erabili den laginaren itzulpen bikote kopurua murriztu egiten da desadostasun kasuak garbitu ondoren. Desadostasunen

garbiketa-prozesuaren ostean, laginaren tamaina 150 sarrera eta 2.616 itzulpen bikotekoa da. Lagin hori erreferentzia eta abiapuntu bezala hartuta, *IC*, *DS* eta bi teknika horien konbinaketaren bidez lortzen diren emaitzak neurtu dira, bataz besteko doitasuna, estaldura eta biak erlazionatzen dituen F-score neurrien bidez. Bestalde, doitasunak edota estaldurak pisu handiagoa duten testuinguruetan teknika bakoitzak duen portaera aztertu ahal izateko F_2 eta $F_{0,5}$ neurrien balioak ere kalkulatu dira. 5.9 taulan teknika bakoitzarekin lortutako emaitza horiek bateraturik ageri dira. Emaitza horietan gurutzaketa gordinaren ostean lorturiko hiztegi gurutzatuaren emaitzak ere gehitu dira, hiztegi honen bidez lorturiko emaitzak *baseline* edo oinarri bezala aurkezten direlarik. Gainontzeko teknikak erabiliz oinarrizko emaitza horiek zenbatean hobetzen diren aztertzea izango da esperimentu honen helburuetako bat.

Hiztegia	Estaldura	Doitasuna	F	$F_{0,5}$	F_2
Baseline *(EU-ZH)*	1,0	0,70	0,82	0,74	0,92
IC *(EU-ZH)*	0,34	0,93	0,49	0,69	0,39
DSF *(EU-ZH)*	0,74	0,73	0,73	0,73	0,74
DSS *(EU-ZH)*	0,35	0,74	0,47	0,61	0,39
IC+DSF *(EU-ZH)*	0,75	0,73	0,74	0,73	0,74

Taula 5.9: Teknika guztien bidez lorturiko emaitzen konparaketa taula

5.9 taulan ikus daitekeen bezala, *baseline* edo oinarri bezala erabiltzen gurutzaketaren ostean lorturiko hiztegiaren emaitzak uste baino hobeak dira, izan ere, itzulpen bikote okerrak kimatzeko teknikarik aplikatu gabe emaitza benetan onak lortzen direla erakusten dute (lan honetan aurkezturiko *EU-ES* esperimentuan baino askoz hobeak). Emaitza altu horiek erreferentzia bezala erabili den laginaren eskuzko ebaluazioren zailtasunaren ondorio zuzenak dira. Izan ere, txinatar hizkuntza ez ezagutzearen ondorioz oso posiblea da hiztegi batean agertu beharko ez luketen itzulpen bikote batzuk ontzat hartu izana: sarreraren antzeko esanahia duten itzulpenak, zuzenak izan arren erabiltzen ez diren itzulpenak... Beste era batera esanda, oso posiblea da erreferentzia bezala erabilitako lagina sortzerakoan malgutasun maila altuegia aplikatu izana. *Baseline* edo oinarri emaitza hain altuak izateak oso zaila bihurtzen du edozein teknika aplikatuz emaitza horiek gainditzea. Emaitzak teknika bakoitzaren ikuspunturik aztertuz, *IC* teknika doitasunaren neurrian bakarrik gailentzen dela nabaria da (0,93), gainerako neurrietan emaitza kaxkarrak lortzen dituelarik, beti ere lortzen duen estaldura-emaitza baxuaren ondorioz (0,34). *DS* teknikaren kasuan, orokorrean atalase malgua aplikatzen den kasuetan *(DSF)* atalase gogorrarekin lortzen diren emaitzak *(DSS)* gainditzen dira, doitasun-emaitzak izan ezik (0,01-eko diferentzia soilik). Beraz, nabarmena da orokorrean atalase malguen erabilpenak emaitza erabilgarriagoak lortzen dituela, lortzen den estaldura maila altua delako. Atalase gogorren aplikazioa doitasun altua beharrezkoa den kasuetan bakarrik da erabilgarriagoa, finkatzen dituen baldintza zorrotzak estalduraren jaitsiera nabarmena suposatzen baitute (0,40-ko beherakada atalase malguekin lorturiko estaldurarekiko). *IC* eta *DS* tekniken konbinaketaren kasuan, bi tekniken emaitzen artean sortzen den dibergentzia mailaren ondorioz hobekuntza minimo bat

lortzen da *DS* teknikarekin atalase malguaren bidez lorturiko emaitzekiko.

Esperimentu honetan aplikatu diren tekniken bidez ez dira *baseline* edo oinarri bezala finkaturiko hiztegiarekin lortutako emaitzak gainditu. Ondorioz, doitasun eta estaldura orokorrari dagokionez badirudi esperimentu konkretu honetan itzulpen okerrak kimatzeko erabiltzen diren teknikek ez dutela ekarpen positiborik egiten. Lehenago aipatu bezala, honen arrazoia laginaren gehiegizko malgutasunean bilatu behar da (hiztegi batean agertu beharko ez luketen itzulpen asko laginean ontzat hartu direlako). Hala ere, kimaketa teknikak automatikoki sorturiko hiztegien beste propietate batzuk hobetzeko erabilgarriak diren aztertu beharra dago. Horrela, aztertu daitekeen beste faktore bat teknika bakoitzaren bidez lortzen diren itzulpenen izaera da, hau da, sarrera bakoitzeko ematen diren itzulpenak probableenak (ohikoenak) edo arraroagoak (ez hain erabiliak edo ez ohikoak) diren. Lagineko sarrera bakoitzaren itzulpen probableenak zein diren identifikatu ahal izateko 3.3.4.2 puntuan aurkezturiko *EU-ZH* corpus paraleloko itzulpen bikoteen maiztasunak erabili dira, sarrera bakoitzeko maizen erabiltzen diren 3 itzulpenak itzulpen probableenak direla suposatuz. Erreferentzia bezala erabiliko diren itzulpen bikoteak lagineko bikoteak izango direnez, itzulpen probableen estaldura da neurtu dena (doitasuna neurtzea zentzugabea litzateke, aztertzen diren bikote guztiak zuzenak izango liratekeelako). *IC* teknikak orokorrean estaldura-emaitza kaxkarra lortzen dituela jakina denez, esperimentu honetan *DS* teknikaren bidez lorturiko hiztegia eta gurutzaketaren ostean lorturiko hiztegia (*baseline*-a) bakarrik aztertzea erabaki da. *DS*-ren kasuan atalase gogorra aplikatu da (*DSS*, edo *TOP3*) eta gurutzaketaren bidez sortutako hiztegiaren kasuan, sarrera bakoitzeko itzulpenen ranking bat sortzeko informaziorik existitzen ez denez, sarrera bakoitzeko hausazko 3 itzulpen aztertzea erabaki da: *R3 (EU-ZH)*. 5.10 taulan, teknika bakoitzarekin lortutako itzulpen probableenen estaldura-emaitzak ikus daitezke.

Hiztegiak	Itzulpen probableenen estaldura
DSS *(EU-ZH)*	0,28
R3 *(EU-ZH)*	0,49

Taula 5.10: Gurutzaketaren eta *DS* teknikaren bidez lorturiko itzulpen probableenen bataz besteko estaldura-emaitzak

5.10 taulan ikus daitekeenez, *DS* teknikarekin sorturiko hiztegia (atalase gogor bat aplikatuz) gurutzaketa gordinaren bidez sorturiko oinarri hiztegia *(baseline)* baino hobea da itzulpen probableenak bakarrik landu nahi diren testuinguruetan. Beraz, nabaria da *DS* teknikak ekarpen handia egin dezakeela itzulpen probableenak lantzen diren testuinguruetan.

Lortutako emaitzetan aztertu daitekeen beste faktore bat aztertu diren sarreren erabilpen maiztasuna da. 5.11 taulan, gurutzaketaren ostean lorturiko hiztegiaren eta teknika bakoitzaren bidez lorturiko hiztegien F-score balioak ageri dira, maiztasun tarte ezberdinen arabera banatuak.

Hiztegia	0<f≤20	20<f<250	f≥250
Baseline *(EU-ZH)*	0,81	0,81	0,80
IC *(EU-ZH)*	0,52	0,51	0,45
DSF *(EU-ZH)*	0,73	0,73	0,74
DSS *(EU-ZH)*	0,5	0,49	0,42
IC+DSF *(EU-ZH)*	0,73	0,74	0,75

Taula 5.11: Teknika guztien bidez lorturiko F-score-en konparaketa taula maiztasun tarteen arabera

5.11 taulan ikusten denez, maiztasun txiki eta ertaineko multzoen artean ez dago diferentzia handiegirik, eta emaitzak orokorrean nahiko konstante mantentzen dira teknika guztien kasuan. Maiztasun handienetan aldiz, baldintza gogorrenak finkatzen dituzten tekniken kasuan emaitzen kalitateak behera egiten duela ikus daiteke. Beheraka hori *IC* eta *DSS*-ren (*DS* teknika atalase gogorra aplikatuz) kasuan nabari da. Bi kasuetan beherakada horren arrazoia maiztasun handieneko sarreren polisemia maila altua da, izan ere, baldintza gogorregiak ezartzeak tekniken bidez lorturiko emaitzen doitasuna hobetu arren estaldura asko kaltetzen du, zaila bihurtzen baita itzulpen posible guztiak behar bezala tratatzea. Estalduraren jaitsiera nabarmen honen ondorioak F-scorean nabaritzen dira.

Bestalde, esperimentu honetan aztertu nahi den beste faktore bat sarreren kategoriaren arabera emaitzen kalitateak duen aldakortasuna da. Izan ere, pentsa daiteke kategoria gramatikal batzuetako sarrerak beste kategoria batzuetakoak baino tratamendu konplexuagoa izan dezaketela, bai orokorrean eta baita teknika konkretu bakoitzaren kasuan ere. Hori aztertzeko, kategoria gramatikalaren arabera teknika ezberdinen bidez sorturiko hiztegi guztien F-scoreak neurtu dira (5.12 taula). Aztertu diren kategoriak izenak, aditzak, adjektiboak eta adberbioak dira.

Hiztegia	Izenak	Aditzak	Adjektiboak	Adberbioak
Baseline *(EU-ZH)*	0,86	0,69	0,84	0,81
IC *(EU-ZH)*	0,56	0,29	0,49	0,46
DSF *(EU-ZH)*	0,78	0,61	0,72	0,72
DSS *(EU-ZH)*	0,55	0,23	0,28	0,51
IC+DSF *(EU-ZH)*	0,78	0,61	0,75	0,73

Taula 5.12: Teknika guztien bidez lorturiko emaitzen konparaketa taula kategoria gramatikalen arabera

5.12 taulako datuen arabera, teknika guztiek orokorrean antzeko portaerak erakusten dituzte. Horrela, izenak gainerako kategoria gramatikaletako sarrerak baino hobeto tratatzen dira. Aldiz, aditz kategoriako sarrerak dira emaitza kaxkarrenak lortzen dituztenak.

5.2.4.2 Sarreren estaldura eta itzulpenen estaldura

HAP masterra

Esperimentu honetan zehar teknika ezberdinak erabiliz sortutako hiztegien beste aspektu aztergarri bat gurutzaketan parte hartzen duten jatorrizko hiztegiekiko duten sarrera eta itzulpenen estaldura da. Neurri honen helburua gurutzaketa-prozesuan parte hartzen duten sarrera eta itzulpen posibleen multzo osotik emaitza hiztegiek tratatzen duten ehunekoa neurtzea da, edo beste era batera esanda, prozesuan zehar ezartzen diren baldintzen ondorioz galtzen den sarrera eta itzulpen kopurua estimatzea. Sarreren nahiz itzulpenen estaldura hori modu automatikoan neurtu dira jatorrizko *EU-EN* eta *EN-ZH* hiztegiak erreferentzia bezala hartuta, izan ere, balio horien kalkuluetan itzulpen bikoteen zuzentasuna ez da kontuan hartzen. 5.13 eta 5.14 tauletan esperimentuan zehar erabilitako teknika ezberdinen bidez sorturiko hiztegien bataz besteko sarreren eta itzulpenen estaldurak ikus daitezke hurrenez hurren.

Teknika	Sarrera kopurua	Estaldura (jatorri hiztegiekiko)
Gurutzaketa gordina	13.544	0,76
IC	3.574	0,20
DSF	9.767	0,55
DSS	9.767	0,55
IC+DSF	10.124	0,57

Taula 5.13: Teknika guztien bidez lorturiko sarreren bataz besteko estalduren konparaketa taula

Teknika	Itzulpen kopurua	Estaldura (jatorri hiztegiekiko)
Gurutzaketa gordina	26.929	0,72
IC	4.607	0,12
DSF	19.592	0,52
DSS	14.638	0,38
IC+DSF	20.102	0,54

Taula 5.14: Teknika guztien bidez lorturiko itzulpenen bataz besteko estalduren konparaketa taula

5.13 eta 5.14 tauletan ageri den bezala, teknika guztien abiapuntua den gurutzaketa aplikatzean dagoeneko jatorrizko hiztegietako hainbat sarrera eta itzulpen emaitzatik kanpo geratzen dira (sarreren %24-a eta itzulpenen %28-a). Gainerako teknikak hiztegi gurutzatu honetatik abiatzen direnez, nabaria teknika bakoitzarekin lor daitekeen bataz besteko sarrera eta itzulpenen estaldura maximoak hiztegi gurutzatuarenak direna. *IC* teknikari dagokionez, bataz besteko sarrera nahiz itzulpenen estaldura oso baxuak dira (sarreren %20-a eta itzulpenen %12-a), teknika honek finkatzen dituen baldintza gogorren ondorioz itzulpen bikote asko kanporatzen baitira. *DS* teknikaren kasuan, sarreren bataz besteko estaldura 0,55-koa da. Itzulpenen estaldura berriz aplikatzen den atalasearen arabera aldatzen da. Ulergarria den bezala, atalase gogorra aplikatzen denean estaldura

baxuagoa da (*DSS:* 0,38) atalase malguagoa ezartzen denean baino (*DSF:* 0,52), ezartzen diren baldintzen gogortasuna handitu ahala itzulpen bikote gehiago geratzen baitira hiztegitik kanpo (*IC* teknikarekin gertatzen den antzera). Azkenik, *IC* eta *DS* tekniken konbinaketa aplikatzen den kasuan bi teknika horiek isolatuan aplikatuz lortzen diren estaldurak gainditzen dira. Hobekuntza honek *IC* eta *DS* tekniken bidez lorturiko sarrera eta itzulpen multzoen artean dibergentzia minimo bat existitzen dela erakusten du, teknika batek harrapatzen ez dituen hainbat sarrera eta bikote besteak harrapatzen dituelarik eta alderantziz.

5.2.5 Esperimentuaren ondorioak

Euskara-Txinera hiztegi elebidunarekin egindako esperimentuetatik hainbat ondorio atera daitezke. Erabili eta aplikatu diren teknikei dagokionez, atera daitezkeen ondorioak Euskara-Gaztelera hiztegiaren esperimentuan atera daitezkeenen oso antzekoak dira. *IC* teknikaren kasuan, garbi geratu da batez ere doitasunak garrantzia duen testuinguruetara egokitzen dela, lortzen dituen estaldura-emaitzak benetan kaxkarrak baitira. Teknikak emaitza erabilgarriak eskain ditzan, prozesuan parte hartzen duten hiztegiek sarrera bakoitzeko ahalik eta itzulpen gehien ematea exijitzen du. Anbiguetate maila altu hori mantentzea oso baldintza gogorra bihurtzen da sarrera askoren kasuan eta ondorioz emaitzen estaldura nahiko kaxkarra izan ohi da. Aldiz, *DS* teknika askoz hobeto egokitzen da edozein testuingurura, aplikatzen den atalasearen aldakortasunari esker. Horrela, doitasunari dagokionez, *IC* teknikak lortzen dituen emaitzak parekatzea lortzen da, eta estaldurari dagokionez *IC* teknikaren bidez lorturiko emaitzak izugarri hobetzen ditu. *DS* teknikaren ahulezia bakarra sarreren maiztasunarekiko duen sentikortasuna da, maiztasun altueneko sarreren kasuan orokorrean emaitzek okerrera egiten dutelarik (*IC* teknikan ez bezala). Hala ere, kontuan eduki behar da *DS* teknika *IC* teknika baino baliabide gehiagotan oinarritzen dela, corpus konparagarri baten erabilera beharrezkoa baita. Hortaz, baliabideen ikuspunturik *IC* teknikak baino baldintza gogorragoak exijitzen ditu. Bi tekniken konbinaketari dagokionez, Euskara-Gaztelera esperimentuan gertatu bezala, *IC* eta *DS* teknikak isolatuan aplikatuz lortzen diren emaitzak gainditzen dira, horrela, eta 4.2.5 puntuan aurreratu bezala, bi teknikek lortzen dituzten emaitzen artean dibergentzia maila minimo bat badagoela frogatuz.

Esperimentu honetan azterturiko beste faktore bat teknika horien aplikazioaren bidez gurutzaketaren ostean lorturiko hiztegiarekiko lorturiko hobekuntza da. Zoritxarrez, eta erreferentzia bezala erabilitako lagina eraikitzerakoan epai irizpide malguegiak aplikatu izanaren ondorioz, hiztegi tradizionaletan agertu beharko ez luketen hainbat itzulpen bikote ontzat hartu dira. Horren ondorioz, erreferentzia bezala erabilitako gurutzaketaren ostean lorturiko hiztegiaren emaitzak espero baino hobeak dira, *IC* eta *DS* teknikak aplikatuz hobetu daitekeen faktore bakarretako bat sarrera bakoitzeko bataz besteko doitasuna delarik. *DS* teknika aplikatuz hobetu daitekeen beste faktore bat itzulpen probableen estaldura da, ohikoenak diren itzulpenak bakarrik ematea beharrezkoa den testuinguru batean (hizkuntza ezezagun batean idatzitako testu baten ulermena adibidez) oinarrizko faktore bihurtzen delarik. Dena den, *IC*, *DS* eta bi tekniken arteko konbinaketak Euskara-Txinera hiztegia-

ren sorkuntzan egin dezaketen ekarpena behar bezala estimatzeko komenigarria litzateke epaietan malgutasun maila murriztuz laginaren sorkuntza-prozesua errepikatzea. Aukera egokiena euskara eta txinera behar bezala ezagutzen dituzten epaileak erabiltzea izango litzateke.

Azkenik, esperimentuetan sorturiko hiztegi baliabidearen erabilgarritasuna aztertu beharra dago. Lorturiko emaitzetan oinarrituz, nabaria da hainbat testuinguruetan erabili ahal izateko zuzenketa lan handia behar duela (kontsulta hiztegi bezala erabiltzeko adibidez). Hala ere, mota honetako baliabideen eskuzko sorkuntzan laguntza bezala erabil daiteke, adibidez sarrera edo itzulpen berriak proposatzeko. Gainera, hizkuntza naturalen prozesamenduaren arloko hainbat atazatan ere erabilgarria izan daiteke (itzulpen automatikoan adibidez). Esperimentu horien hurrengo urratsetako bat sorturiko hiztegiak testuinguru horietan duen erabilgarritasuna frogatzea izango litzateke, proba eta testak burutuz.

6 Ondorioak

Lan honetan hiztegi elebidunak automatikoki sortzea ahalbidetzen duten tekniken azterketa sakon bat burutu da, horretarako aukeratu den testuingurua baliabide gutxien duten hizkuntzena delarik. Lan honen helburu nagusiak bi dira, bata pibotaje teknika ezberdinek testuinguru honetan duten portaera aztertzea eta bestea teknika horien bidez sortzen diren hiztegien erabilgarritasuna eta kalitatea neurtzea. Kasu guztietan, sorturiko hiztegien erabilpen testuinguruaren faktorea kontuan hartu beharra dago, erabilera ezberdin bakoitzean sortzen diren beharrak ez baitira berdinak.

Pibotaje teknikei dagokionez, lan honetan baliabide gutxien duten hizkuntzen baldintzetara ondoen egokitzen diren bi teknikak aztertu dira: *IC* teknika eta *DS* teknika. Egindako esperimentu guztiek teknika horiek oso izaera ezberdina dutela erakusten dute, bakoitza paradigma eta baliabide ezberdinetan oinarritzen baita. *IC* teknikaren kasuan, ezartzen diren baldintzak nahiko zorrotzak dira, eta abiapuntu bezala erabiltzen diren hiztegietako sarrerek ahalik eta itzulpen gehien eskaintzea ezartzen du beharrezko baldintzatzat. Baldintza gogor horien ondorioz, lortzen diren emaitzen propietate nagusia doitasun maila altua da. Zoritxarrez, baldintza gogor horien albo kalteak estalduran islatzen dira, emaitza bezala doitasun handiko hiztegi txikiak lortzen direlarik. Aldiz, *DS* teknika askoz ere malguagoa da, sarrera bakoitzaren itzulpenak identifikatzeko aplikatzen duen antzekotasun neurriaren exijentzia maila egoera ezberdinen arabera egokitu baitaiteke. Horrela, *DS* teknika *IC* teknika bezain erabilgarria da doitasun altua beharrezkoa den inguruneta, baina baita estaldurak garrantzia handiagoa hartzen duen testuinguruetan ere (*IC* teknika ez bezala). Dena den, *IC* teknikak baino baliabide gehiago erabiltzen ditu (hiztegi elebidunez gain, corpus konparagarriak ere beharrezkoak dira), eta beraz, nahiz eta metodologia aldetik malguagoa izan, baliabideen ikuspuntutik *DS* teknika *IC* teknika baino zorrotzagoa da. Ondorioz, garbi dago teknika bakoitzak bere alde onak eta txarrak dituela. Beraz, bata edo bestearen arteko aukeraketa egiteko eskura dauden baliabideak eta lortu nahi den hiztegi mota eduki behar dira kontuan. Beharrezko baliabide guztiak eskura izanda, *DS* teknikaren emaitzak orekatuagoak dira, aldiz, baliabideak mugatuagoak diren kasuetan *IC* teknikarekin emaitza onargarriak lor daitezke. Bestalde, bi teknikak aplikatu ahal izateko baldintzak betetzen diren kasuetan, bi tekniken arteko konbinaketaren bidez teknika bakoitzaren emaitzak gainditu daitezke, paradigma eta baliabide ezberdinetan oinarritzearen ondorioz bi tekniken emaitzetan dibergentzia kasuak aurki baitaitezke.

Lan honen bigarren helburu nagusia pibotaje tekniken bidez sortzen diren hiztegien erabilgarritasuna aztertzea da. Esperimentuen emaitzen arabera, nabaria da lortzen diren hiztegiak ez direla publikatzeko moduko hiztegiak, eskaintzen dituzten itzulpenen zati handi bat zarata baita. Hala ere, beste zenbait erabilpen esparrutan duten erabilgarritasuna aztertu beharra dago. Izan ere, hainbat esparru edo testuingurutan ezartzen diren baldintzak publikatzen diren hiztegienak baino malguagoak dira. Horrela, automatikoki sortutako hiztegi horiek oso lagungarriak izan daitezke, adibidez, dagoeneko existitzen diren hiztegien estaldura hobetzeko, horretarako hiztegi automatikoak eskaintzen dituen itzulpen proposamenak eskuz balidatzen badira. Bestalde, hizkuntza naturalen prozesamenduaren arloko hainbat atazetan ere hiztegi horiek erabilgarriak izan daitezke (itzulpen

automatikoan esaterako), mota honetako atazatan erabiltzen diren baliabideek bete beharreko baldintzak malguagoak izaten baitira.

Bukatzeko, garrantzitsua da lan honetan aurkeztu diren esperimentuen hurrengo urrats posibleak identifikatu eta definitzea. Hasteko, beharrezkoa da lan honetan zehar burutu diren esperimentuak hizkuntza konbinazio gehiagorekin burutzea. Izan ere, esperimentu horiek zenbat eta hizkuntza konbinazio gehiagoren gainean burutu orduan eta ondorio sendoagoak lortuko dira. Bestalde, lan honetan lorturiko hiztegien ebaluazioa modu isolatuan burutu da, hau da, hiztegiaren izaera bakarrik aztertu da (ebaluazio intrintsekoa). Hiztegi horien erabilgarritasuna sakonago aztertzeko, interesgarria litzateke hiztegi horien erabilerak ataza zehatz batean duen eragina aztertzea (ebaluazio estrintsekoa). Adibidez, itzulpen automatikoko ataza batean hiztegi horien erabilpenak emaitzetan duen eragina aztertu daiteke. Mota honetako azterketa batek hiztegi elebidun horien erabilgarritasuna hobeto estimatzeko balioko luke.

Erreferentziak

E. Agirre, I. Alegria, G. Rigau, eta P. Vossen. Mcr for clir. In *Procesamiento del lenguaje natural 38*, pages 3–15, Edinburgh, Scotland, 2007.

F. Bond eta K. Ogura. Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary. *Language Resources and Evaluation*, 42(2), 2007. ISSN 1574-020X. doi: 10.1007/s10579-007-9038-4.

F. Bond, R. Sulong, T. Yamazaki, eta K. Ogura. Design and construction of a machine-tractable Japanese-Malay dictionary. *Proceedings of ASIALEX, SEOUL*, 2001(2001): 200–205, 2001.

M. Erdmann, K.Ñakayama, T. Hara, eta S.Ñishio. An approach for extracting bilingual terminology from wikipedia. In *Proceedings of the 13th international conference on Database systems for advanced applications*, DASFAA'08, page 380–392, Berlin, Heidelberg, 2008a. Springer-Verlag. ISBN 3-540-78567-1, 978-3-540-78567-5. ACM ID: 1802552.

M. Erdmann, K.Ñakayama, T. Hara, eta S.Ñishio. A bilingual dictionary extracted from the wikipedia link structure. In *Proceedings of the 13th international conference on Database systems for advanced applications*, DASFAA'08, page 686–689, Berlin, Heidelberg, 2008b. Springer-Verlag. ISBN 3-540-78567-1, 978-3-540-78567-5. ACM ID: 1802593.

N. Ezeiza, I. Aduriz, I. Alegria, J.M. Arriola, eta R. Urizar. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *COLING-ACL'98*, pages 380–384, Edinburgh, Scotland, 1998.

P. Fung. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the Third Workshop on Very Large Corpora*, page 173–183, Edinburgh, Scotland, 1995. Somerset, New Jersey. Association for Computational Linguistics.

P. Gamallo eta J. Pichel. Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora. In *Computational Linguistics and Intelligent Text Processing*, page 473–483, Iasi, Romania, 2010. Springer.

V. István eta Y. Shoichi. Bilingual dictionary generation for low-resourced language pairs. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, page 862–870, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-62-6. ACM ID: 1699625.

H. Kaji eta Y. Morimoto. Unsupervised word sense disambiguation using bilingual comparable corpora. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, page 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1072228.1072286. ACM ID: 1072286.

H. Kaji, S. Tamamura, eta D. Erdenebat. Automatic construction of a Japanese-Chinese dictionary via english. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008. ISSN 2-9517408-4-0.

G. Miller, R. Beckwith, C. Fellbaum, D. Gross, eta K. Miller. Introduction to wordnet: An online lexical database. In *Int J Lexicography 3(4)*, pages 235–244, Edinburgh, Scotland, 1990.

K. Paik, S. Shirai, eta H.Ñakaiwa. Automatic construction of a transfer dictionary considering directionality. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, MLR '04, page 31–38, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. ACM ID: 1706243.

W. Peters, P. Vossen, P. Díez-Orzas, eta G. Adriaens. Cross-linguistic alignment of wordnets with an inter-lingual-index. In *Computers and the Humanities 32*, pages 221–251, Edinburgh, Scotland, 1998.

E. Prochasson, E. Morin, eta K. Kageura. Anchor points for bilingual lexicon extraction from small comparable corpora. In *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, pages 284–291, Ottawa, Canada, August 2009.

R. Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, page 519–526, Edinburgh, Scotland, 1999. College Park, USA. ACL.

G. Rohit eta M. Tandon. An iterative approach to extract dictionaries from wikipedia for under-resourced languages. In *ICON 2010*, Edinburgh, Scotland, 2010. IIT Kharagpur, India.

M. Sammer eta S. Soderland. Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. In *Proceedings of the MT Summit XI*, pages 399–406, 2007.

D. Shezaf eta A. Rappoport. Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, page 98–107, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ACM ID: 1858692.

S. Shirai eta K. Yamamoto. Linking english words in two bilingual dictionaries to generate another language pair dictionary. In *Proceedings of ICCPOL*, page 174–179, 2001.

S. Shirai, K. Yamamoto, eta K. Paik. Overlapping constraints of two step selection to generate a transfer dictionary. In *ICSP-2001*, pages 731–736, Daejeon, South Korea, 2001. doi: 10.1.1.20.7639.

J. Sjöbergh. Creating a free digital Japanese-Swedish lexicon. In *Proceedings of PACLING 2005*, 2005.

K. Tanaka eta K. Umemura. Construction of a bilingual dictionary intermediated by a third language. *Proceedings of the 16th international conference on computational linguistics (COLING'94)*, 39:297–303, 1994.

H. Tseng. A conditional random field word segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing*, Edinburgh, Scotland, 2005.

M. Tsuchiya, A. Purwarianti, T. Wakita, eta S.Ñakagawa. Expanding Indonesian-Japanese small translation dictionary using a pivot language. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, page 197–200, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. ACM ID: 1557826.

T. Tsunakawa, N. Okazaki, eta J. Tsujii. Building bilingual lexicons using lexical translation probabilities via pivot languages. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

P. Vossen. Introduction to eurowordnet. In *Computers and the Humanities 32*, pages 73–89, Edinburgh, Scotland, 1998. Special Issue on EuroWordNet.

K. Yu eta J. Tsujii. Bilingual dictionary extraction from wikipedia. In *Proceeding of Machine Translation Summit XII*, Edinburgh, Scotland, 2009.