EUSKAL HERRIKO UNIBERTSITATEA (UPV/EHU)
Euskal Hizkuntza eta Komunikazioa Saila

Doktorego-tesia

Euskarazko egitura sintaktiko konplexuen analisirako eta testuen sinplifikazio automatikorako proposamena

Itziar Gonzalez Dios

2016

EUSKAL HERRIKO UNIBERTSITATEA (UPV/EHU)
Euskal Hizkuntza eta Komunikazioa Saila

Euskarazko egitura sintaktiko konplexuen analisirako eta testuen sinplifikazio automatikorako proposamena

Itziar Gonzalez Diosek Arantza Díaz de Ilarrazaren eta María Jesús Aranzaberen zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Doktore titulua eskuratzeko aurkeztua.

Donostia (2016).

Caminante,
no hay camino,
se hace camino al andar.

Antonio Machado

Eskerrak

Lehenik eta behin eskerrak eman nahi nizkieke tesi-lan honen zuzendariak izan diren Arantzari eta Maxuxi, zuen denboragatik, arretagatik eta konfiantzagatik, bidea irekitzeagatik eta bidean gidatzeagatik. Eskerrak eman nahi dizkiet Xabier Artolari eta Izaskun Aldezabali azken orrazketak eta zuzenketak egiteagatik.

Modu batera edo bestela, lan honetan kolaboratu duzuenoi: Itziar Aduriz, Rodrigo Agerri, Manex Agirrezabal, Edurne Aldasoro, Itziar Aldabe, Izaskun Aldezabal, Begoña Altuna, Nora Aranberri, Maxux Aranzabe, Zuhaitz Beloki, Klara Ceberio, Ainara Estarrona, Uxoa Iñurrieta, Mikel Iruskieta, Gorka Labaka, Mikel Lersundi, Oier Lopez de Lacalle, Inigo Lopez-Gazpio, Unai Lopez-Novoa, Vanessa Martin, Itziar Otaduy, Arantxa Otegi, Iñaki San Vicente, Aitor Soroa eta Larraitz Uria (esango nuke denak zaudetela... norbait ahaztu bazait, barkatu!). Zuen " marroitxoak" nire fruitu izan direlako... Amaiari, Estherri eta Kikeri beti laguntzeko prest egon zaretelako!

Nola ez, Ixa taldeko kideei kafe orduak, bazkaltzeko tarteak eta "bestelako ekintzak" ezinbestekoak izan dira tesia aurrera eramateko eta arinagoa izateko! Eta batez ere, bulegokidei eta bulegokide izandakoei. Gora 314 bulegoa eta krisi komiteak! Bestelako "lantxoetan" izan ditudan taldekideei, zuen esperientzia irakasle onena izan baita!

3. pisuan pasillotik bueltaka aurkitu ditudanei, eta, bereziki, Mendiri,

txiste txarrekin irribarrea ateratzeagatik... babesagatik eta laguntzagatik!

Il gruppo ItaliaNLP Lab, per la vostra hospitalità, e farmi sentire come a casa! E Chiara, la megliore coinquilina! Non ho parole, grazie mille a tutti!!!

Koadrilakoei eta uniko lagunei, batzuk urrun besteak gertu baina hor egon zeatelako!

Y, por último, a los de casa, por ser vosotros, por estar siempre a mi lado y dispuestos a todo por ayudarme. Eskerrik asko, aita eta ama! Esti, eredua izategatik, eta Izar, egunak alaitzeagatik! A las abuelas, porque una sonrisa vuestra vale más que nada, y a los abuelos, donde quiera que esteis, porque espero que estéis orgullosos de mi!

Bidean topatutako edonori, bidea erraztu edo zaildu duen orori, indartsuago eta ulerkorragoa bihurtzeagatik!

Mila esker!!!

Esker instituzionalak

Eusko Jaurlaritzako Hezkuntza, Unibertsitate eta Ikerketa Sailari, ikerketalan hau egiteko emandako ikertzaileak prestatzeko bekarengatik (BFI-2011-392).

Laburpena

Tesi-lan honetan, euskarazko testuen konplexutasuna eta sinplifikazioa automatikoki aztertzeko lehen urratsak egin ditugu. Testuen konplexutasuna aztertzeko, testuen sinplifikazio automatikoa helburu duten beste hizkuntzetako lanetan eta euskarazko corpusetan egindako azterketa linguistikoan oinarritu gara. Azterketa horietatik testuak automatikoki sinplifikatzeko oinarri linguistikoak ezarri ditugu. Konplexutasuna automatikoki analizatzeko, ezaugarri linguistikoetan eta ikasketa automatikoko tekniketan oinarrituta ErreXail sistema sortu eta inplementatu dugu.

Horretaz gain, testuak automatikoki sinplifikatuko dituen Euskarazko Testuen Sinplifikatzailea (EuTS) sistemaren arkitektura diseinatu dugu, sistemaren modulu bakoitzean egingo diren eragiketak definituz eta, kasu-azterketa bezala, informazio biografikoa duten egitura parentetikoak sinplifikatuko dituen Biografix tresna eleaniztuna inplementatuz.

Amaitzeko, Euskarazko Testu Sinplifikatuen Corpusa (ETSC) osatu dugu. Corpus hau baliatu dugu gure sinplifikaziorako azterketetatik ateratako hurbilpena beste batzuekin erkatzeko. Konparazio horiek egiteko, etiketatze-eskema bat ere definitu dugu.

"Nazioarteko Doktorea" aipamena lortzeko Euskal Herriko Unibertsitatearen eskakizunei jarraituz, tesi-txosten honen ingelesezko bertsio laburtua ondorengo helbide honetan aurki daiteke:
http://ixa.eus/Ixa/Argitalpenak/Tesiak/1462951351/publikoak/english_report.pdf

Abstract

In this thesis we have paved the way for the automatic readability assessment and text simplification in the automatic processing of Basque. In order to analyse the complexity of the texts, we have considered the works for other languages targeted to automatic text simplification and we have performed linguistic analyses in Basque corpora. Based on these analysis we have set the linguistic foundations to simplify texts automatically. To assess the readability of the texts automatically, we have implemented ErreXail, a system based on linguistic features that uses machine learning techniques. To simplify texts automatically, we have defined the operations that the text simplification system EuTS should perform and we have connected them with the modules of the architecture. We have also provided the linguistic information these modules need. As case study, we have implemented a multilingual tool that simplifies parenthetical structures containing biographical information, and we have shown that the results of the linguistic analysis for Basque are also useful for other languages. To contrast our corpus-study-based approach, we have created the ETSC-CBST corpus that contains original and simplified texts. To make the comparison among different approaches, we have defined an annotation scheme.

The English summary of this report can be found at
http://ixa.eus/Ixa/Argitalpenak/Tesiak/1462951351/publikoak/english_report.pdf

Gaien aurkibidea

Irudien zerrenda

Taulen zerrenda

1 Tesi-lanaren aurkezpen orokorra

1.1 Sarrera

Informazioaren gizartean milaka testu ekoizten dira egunero, baina testu horiek guztiak ez dira jende guztiarentzat eskuragarriak, konplexuak direlako prozesatzeko. Idatzizko testuen[1] konplexutasuna, egun, pertsona askorentzat arazo larria da, bai ikasketa-maila egokia ez dutelako bai bestelako arazoak dituztelako. Pertsonentzat ez ezik, Hizkuntzaren Prozesamenduko (HP) aplikazioentzat ere arazo bihurtu da. Horren adibide dugu, esaterako, 1.1 taulako euskarazko esaldi baten ingeleseko itzulpen automatikoa[2].

Esaldia	**Esaldiaren itzulpena**
1962an Charles De Gaulle eta Konrad Adenauer Bonnen elkartu zirenean 55 miloi lagun bizi ziren herrialde horretan, eta 47 milioi Frantzian.	Charles De Gaulle and Konrad Adenauer in Bonn, when 55 million people were living together in this country, and 47 million in France.

1.1 taula – Google Translate itzultzaile automatikoarekin egindako ingelesezko esaldi baten itzulpena

Google Translate erabilita itzuli den esaldi horretan, jatorrizko euskarazko esaldiaren mendeko perpauseko "elkartu" aditzaren ordaintzat perpaus nagusiko *together* hitza eman du ingelesez. Ondorioz, itzulitako esaldiaren perpaus

[1] Txosten honetan hemendik aurrera *testu* hitza erabiltzen dugunean, idatzizko testuei buruz ariko gara.

[2] Esaldiaren itzulpen automatikoa `https://translate.google.es/` web-zerbitzua erabiliz egin dugu, 2013ko abenduan.

nagusiko aditza falta denez, ez dakigu zer egiten zuten Charles De Gaullek eta Konrad Adenauerek Bonnen. Gainera, perpaus nagusia zena mendeko perpaus bihurtu du itzulpenean, *when* hitzarekin hasten baita. Ikusten dugunez, itzulpenean euskarazko esaldiaren perpausen elementuak nahastu egin ditu. Nahaste hori, hain zuzen ere, perpausen konplexutasunarekin lotuta dago, mendeko perpausak dituzten esaldiak itzultzean errore-kopurua handitu egiten baita.

Testuen eta esaldien konplexutasunak sortzen dituen arazo horiei aurre egin nahi izan diegu tesi-lan honetan eta, horretarako, HPko bi ikerketa-lerro aztertu ditugu: Testuen Sinplifikazio Automatikoa (TSA), ingelesez *Automatic Text Simplification* (ATS), eta Testuen Konplexutasunaren Analisia (TKA), ingelesez *Readability Assessment* (RA). TSAk testu konplexuak sinpleago bihurtzea du helburu, betiere jatorrizko esanahiari eutsiz; TKAk, berriz, testu bat konplexua den ala ez edo zein konplexutasun-maila duen aztertzen du. Bi ikerketa-lerro horiek berebiziko garrantzia dute aro digitalean; izan ere, gure gizartean ekoizten diren testuen konplexutasuna eskuz aztertzea eta testu horiek sinplifikatzea ataza garestia eta motela da. HPko teknologiak baliatuta, ordea, lan horiek erraztu eta azkartu egiten dira.

TSA Hizkuntzaren Sorkuntzaren (HS) arloan, ingelesez *Natural Language Generation* (NLG), kokatu izan da, testuak sinplifikatzean testu horietan hizkuntza sortzen delako. TSAn bi motatako sinplifikazioak egiten dira nabarmenki: sinplifikazio sintaktikoa eta sinplifikazio lexikala. Sinplifikazio sintaktikoan, egitura sintaktikoak berridatzi egiten dira egitura sinpleagoak lortuz, eta sinplifikazio lexikalean, hitz zailak edo maiztasun gutxikoak ezagunagoekin ordezkatzen dira jatorrizko testuaren esanahiari eutsiz. Sinplifikazio horiek egiteko sistemak erregeletan oinarritutakoak, datuetan oinarritutakoak edo hibridoak dira.

Horretaz gain, testuak sinplifikatzean, normalean, testu horiek nork jasoko dituen (helburu-taldea) kontuan izaten da. Oro har, bi helburu-talde nagusi daude: gizakiak eta makinak. Jarraian, testu sinplifikatuen baliagarritasuna talde bakoitzean azalduko dugu.

- **Gizakiak:** atzerriko hizkuntzen ikasleak, alfabetitazio-arazoak dituztenak, haurrak, gaixotasunak dituztenak...
  - Atzerriko hizkuntzak ikasten ari direnak: ikasketa-prozesu horretan, egiturak mailaka ikasten direnez, ez dituzte egitura guztiak ezagutzen ikasketa-prozesua aurreratua izan arte.

  - Analfabeto edo gutxi alfabetatuak: ikasketa/irakurketa-maila baxua dutenez, testu landuak zailak iruditzen zaizkie.
  - Haurrak: oraindik hizkuntza osoa barneratuta ez dutenez, kontzeptu batzuk ez dituzte ulertzen. Kasu honetan, sinplifikazioa batez ere lexiko-mailakoa da.
  - Afasikoak: hizkuntza ulertzeko eta hitz egiteko gaitasuna galdu dutenez, egitura jakin batzuk ulertzeko gaitasuna ere galdu dute.
  - Gorrak: munduaren kontzeptualizazio ezberdina dutenez, hizkuntza arrunta ulertzea zaila egiten zaie.
  - Gaixotasun kognitiboak dituztenak: Alzheimerraren eta antzeko beste gaixotasunen ondorioz, ulertzeko gaitasuna galtzen dutelako.

- **Makinak:** HPko sistemak zein bestelako gailuak
  - HPko sistema aurreratuak[3] (analizatzaile sintaktikoak [*parser*]), itzulpen automatikoa, galde-erantzun sistemak eta laburpen automatikoak egiteko sistemak): aurreprozesu moduan erabil daiteke sinplifikazioa, esaldi laburrak errazago eta hobeto prozesatzen baitira.
  - Pantaila txikiak dituzten gailuak (sakelako telefonoak, PDAk, tabletak...): esaldi laburrekin informazioa modu erosoagoan agertzen da.

Horiek guztiak izanik, TSA HPan oihartzun handia izan duen ikerketalerroa da. Horren lekuko dira azkeneko urteetan HPko konferentzia nagusietan (*Language Resources and Evaluation Conference* [LREC], *European Chapter of the Association for Computational Linguistics* [EACL] eta *International Conference on Computational Linguistics* [Coling]) antolatu diren *workshop*ak: *Predicting and Improving Text Readability for target reader populations* (PITR) 2012an, 2013an eta 2014an, *Natural Language Processing for Improving Textual Accessibility* (NLP4ITA) 2012an eta 2013an, eta *Automatic Text Simplification - Methods and Applications in the Multilingual Society* (ATS-MA) 2014an. 2016an TSArekin lotutako beste bi *workshop*

[3]HPko tresnak helburu dituzten lanetan, sinplifikazioa adierazteko esaldien sinplifikazioa (*sentence simplification*) terminoa erabili izan da hasierako lanetan. Azken urteetan, berriz, termino hori ez da hainbeste erabili.

antolatu dira: *Improving Social Inclusion using NLP: Tools and resources* (ISI-NLP) eta *Quality Assessment for Text Simplification* (QATS). Azken horren barnean txapelketa (*shared task*) bat ere antolatu da.

Workshop horietan aurkeztu diren lanetan, eta horietaz gain, beste konferentzietan eta aldizkarietan argitaratu direnetan ere, testuen sinplifikazio automatikoa hainbat modutan tratatu da. Testuak automatikoki sinplifikatzeko hurbilpenen artean lehenengo bereizketa sortzen duena informazioaren kudeaketa da. Lan batzuetan jatorrizko testuetan dagoen informazio guztiari eusten zaio (Siddharthan, 2006; Gasperin *et al.*, 2009a; Aranzabe *et al.*, 2012a); beste lan batzuetan, aldiz, beharrezkoa ez den informazioa ezabatu egiten da (Bott *et al.*, 2012b; Barlacchi eta Tonelli, 2013). Gure ustez, bigarren hurbilpen horiek laburpen automatikoen alorreko ikuspuntua hartzen dute, sinplifikatzeaz gain testuak laburtu egiten dituztelako. Hain zuzen ere, TSA HPko beste ikerketa-lerro batzuekin nahastu izan da, ataza horien antzekotasuna dela kausa. Horien artean *automatic summarization* edo laburpen automatikoen ikerketa-lerroa eta horrekin lotuta dauden *sentence reduction*, *sentence compression* edo *sentence fusion* (esaldien fusioa edo trinkotzea) atazak aurki ditzakegu. Horietan ez bezala, sinplifikazioan lan gehienetan behintzat, informazio guztia mantentzeko joera dago. TSArekin lotura zuzena duten beste ataza bat parafrasiak ikastea eta sortzea (*paraphrase acquisition and generation*) da, testuak sinplifikatzeko maiz parafrasiak erabili eta berridazketak egiten direlako.

Informazioaren kudeaketaz gain, badira sinplifikazioaren izenean egiten diren beste hurbilpen eta ikuspuntu ezberdinak ere. Adibidez, FIRST proiektuko[4] lanetan, testuak norbanakoari egokitzeko tresnak sortu dituzte. Proiektuaren koordinatzaileak beren lana sinplifikazio izenarekin baino "testuen pertsonalizazio" izenarekin bataiatu zuen (ahozko komunikazioa, ATS-MA *workshop*a 2014). Izan ere, norbanako bakoitzari zuzenduriko testu-egokitzapena da beren lanetan egiten dutena.

2014ko ATS-MA *workshop*ean aurkeztutako beste bi lanek ere ikuspuntu berriak gehitzen dizkiete orain arte literaturan aurkitutako lanei. Alde batetik, testu historikoak gaurko grafiara ekartzea (Vertan eta von Hahn, 2014) izan da sinplifikazio ikuspuntu berri bat, eta bestetik patenteak biltzen dituzten dokumentuak grafo bidez ematea eta erlazioak jartzea (Sheremetyeva, 2014). Testuak gaurko grafiara ekartzea, gure ustez, testu-normalizazioa da eta ez sinplifikazioa. Ulertzen dugu testua gaur egungo grafiara ekartzean

[4] http://www.first-asd.eu/ (2016ko urtarrilean atzituta)

testua errazten dela edo irakurterrazagoa egiten dela (*legibility*), baina ikuspuntu hori ez dator bat guk egingo dugun sinplifikazioarekin. Dokumentuak grafo bihurtzeari eta erlazioak jartzeari dagokionez, ulertzen dugu testuaren interpretazioa errazagoa izan daitekeela egitura berriarekin, baina aurreko ikuspuntuarekin bezalaxe ez dator bat guk dugun ikuspegiarekin.

Sinplifikazioarekin elkarrekintza zuzena duen beste alor bat hizkuntza kontrolatuena (*controlled languages*) da. Larrialdien domeinuko testuak sinplifikatzeko, adibidez, hizkuntza kontrolatuak erabiltzea proposatu da (Temnikova, 2012). Irakurketa errazak (*plain language*), bestetik, badu zerikusi zuzena. Erakundeek ematen dituzten gidalerro horietan oinarrituz TSArako hainbat lan egin dira (Bott *et al.*, 2012b; Mitkov eta Štajner, 2014). Beraz, paragrafo honetan aipatu ditugun bi kontzeptu horiek sinplifikazioa bideratzeko metodoak dira.

Euskarari dagokionez, guk dakigula ez dago euskararako hizkuntza kontrolaturik HPan eta irakurketa errazaren presentzia[5] oso berria da. Dena den, aipatu nahi dugu 1978an Imanol Berriatua euskaltzainak "oinarrizko euskara" ikasteko metodoa plazaratu zuela Israelgo esperientzian[6] oinarrituta.

TSAn jarraitzen diren hurbilpenak 1.2 taulan jaso ditugu.

Lana	**Informazioa mantenduz**	**Informazioa kenduz**	**Irakurketa erraza**	**Hizkuntza kontrolatuak**	**Bestelakoak**
Siddharthan (2006)	✔	-	-	-	-
Gasperin *et al.* (2009a)	✔	-	-	-	-
Temnikova (2012)	-	-	-	✔	-
Bott *et al.* (2012b)	-	✔	✔	-	-
Barlacchi eta Tonelli (2013)	-	✔	-	-	-
Vertan eta von Hahn (2014)	-	-	-	-	✔

(Jarraipena hurrengo orrialdean)

[5]Euskarazko irakurketa errazeko lehenengo liburuak 2014an argitaratu zituen Gaumin argitaletxeak.

[6]Israelen hebreera berreskuratzeko hebreera errazaren metodoak ikasi zituen.

Lana	Informazioa mantenduz	Informazioa kenduz	Irakurketa erraza	Hizkuntza kontrola-tuak	Bestelakoak
Sheremetyeva (2014)	-	-	-	-	✔

1.2 taula – TSA hainbat lan hurbilpenen arabera

Testuen konplexutasunaren analisian, bestalde, testuak sinpleak edo konplexuak diren, edo zein mailari dagozkion aztertzen da. Horretarako, testuen ezaugarri linguistikoak edo/eta estatistikoak kontuan hartzen dira. Analisi horretan testuaren edukia analizatzen da (*readability*) eta ez testu horren itxura (*legibility*), hau da, testuaren letra-tamainak eta -motak, justifikazioa, espazioak eta abar ez dira analizatzen. Tesi-lan honetan, batez ere, TSAri zuzendutako TKA aztertuko dugu. TKA sistema horien helburua testuak sinpleak ala konplexuak diren adieraztea da sinplifikatu behar diren ala ez jakiteko. TKA TSAren ebaluazio bezala ere erabili da.

Tesi-lan honetan, beraz, TSA eta TKA euskarara ekartzeko ahalegina egingo dugu, atzerriko hizkuntzetan egindako lanak aztertuz.

1.2 Tesi-lanaren motibazioa eta helburuak

Tesi-lan hau Euskal Herriko Unibertsitateko diziplina arteko Ixa ikerketa-taldean[7] garatu dugu. Ixa taldeak 27 urte daramatza euskararen tratamendu automatikoan, oinarrizko ikerketa ez ezik, baliabideak eta aplikazio aurreratuak ere sortzen. Gure lana aplikazio aurreratuen multzoan kokatzen da batik bat, baina egindako ikerketak HPan ezinbestekoa den hizkuntzaren formalizazio konputazionaletik ere asko du.

Tesi-lan honen motibazio nagusia testu sinpleagoak edo errazagoak lortzea da HPko tresna aurreratuetan esaldi luzeek eta konplexuek sortzen dituzten arazoak ebazteko eta euskara ikasten ari diren pertsonei laguntzeko. Izan ere, testu konplexutzat hartuko diren testuak sinpleago bihurtzeko sistema diseinatu nahi dugu. Hori gauzatzeko, Ixa ikerketa-taldearen tresnen eta baliabideen berrerabilpena aztertu eta kontuan hartuko dugu.

Motibazio horiekin bi helburu ditugu: 1) alde linguistikotik, konplexutasuna aztertzea eta sinplifikazio-proposamenak egitea eta 2) tratamendu

[7] http://ixa.eus (2016ko urtarrilean atzituta)

konputazionaletik, testuen konplexutasuna neurtuko eta testuak automatikoki sinplifikatuko dituzten sistemak informazio linguistikoarekin hornitzea eta inplementatzea. Helburu horiek lortzeko, erantzuna emango diogu lau multzo hauetako ikerketa-galderei.

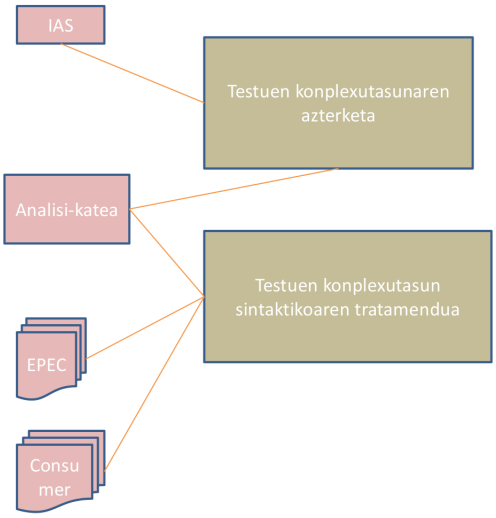
- **Konplexutasunaren azterketa:** Zer da konplexutasuna? Nola kalkula daiteke? Zein dira euskarazko egitura konplexuak?
- **Konplexutasunaren tratamendua (sinplifikazioa):** Egitura konplexuak nola sinplifika daitezke? Zein da egitura horiek sinplifikatzeko prozesua? Zein eragiketa egin behar dira?
- **Baliabideak:** Zein baliabide behar dira prozesu hori aurrera eramateko?
- **Beste hizkuntzekiko konparazioa:** Zein behar berezi ditu euskarak erdal hizkuntzekin alderatuta?

Gure lanaren abiapuntuan izan ditugun baliabideak eta tresnak konplexutasunaren azterketaren edo/eta konplexutasun sintaktikoaren tratamenduaren arabera jaso ditugu 1.1 irudian. Baliabide eta tresna horiek Idazlanen Autoebaluaziorako Sistema (IAS) (Castro-Castro *et al.*, 2008), Ixa ikerketa-taldearen analisi-katea (Aduriz *et al.*, 2004), eta *Euskararen Prozesamendurako Erreferentzia Corpusa* (EPEC) (Aduriz *et al.*, 2006a) eta *Consumer* corpusa (Alcázar, 2005) dira. Horien berrerabilpena aztertu ondore iruditu zaigu Ixa taldearen analisi-katea testuak automatikoki analizatzeko eta IAS konplexutasuna automatikoki analizatzeko erabilgarriak direla, eta EPEC eta *Consumer* corpusak, berriz, konplexutasunaren eskuzko azterketa egiteko.

1.3 Tesi-txostenaren eskema

Tesi-lanaren aurkezpen orokor horren ondoren, txosten hau lau atal nagusitan banatu dugu: TSAn eta TKAn egin diren erdaretako lanen azterketa, azterketa linguistikoa, tratamendu konputazionala eta eskuz sinplifikatutako testuen analisia. Atal horietako bakoitza kapitulu hauetan azaldu dugu:

- Gure lanaren aurrekariak (erdaretako lanen azterketa) aurkeztuko ditugu 2. kapituluan.



1.1 irudia – Tesi-proiektuaren hasieran genituen baliabideak eta tresnak

- Azterketa linguistikoan, egitura sintaktiko konplexuen analisi linguistikoa egingo dugu. Azterketa horren emaitzak 3. kapituluan emango ditugu.

- Tratamendu konputazionalean egin ditugun lanen azalpena 4. kapituluan hasiko dugu, euskarazko TSArako gure hurbilpena eta TSA egiteko tresnak azalduz. 5. kapituluan konplexutasunaren analisi automatikoa egiten duen ErreXail sistema aurkeztuko dugu, eta 6. kapituluan testuen sinplifikazio automatikoa egingo duen EuTS sistemaren diseinua azalduko dugu.

- Eskuz sinplifikatutako testuen analisiaren atalean, 7. kapituluan *Euskarazko Testu Sinplifikatuen Corpusa* (ETSC) ezagutaraziko dugu.

Ondorioekin eta etorkizuneko lanekin emango diogu amaiera tesi-txostenari, 8. kapituluan. Kapitulu horietaz gain, hiru eranskin daude tesi-txosten honetan: perpaus adberbialen egiturak (A eranskina), egitura konplexuak sinplifikatzeko erregelak (B eranskina) eta ETSC corpusa zabaltzeko eragiketen zerrenda (C eranskina).

1.4 Argitalpenak eta sariak

Sarrera-kapitulua amaitzeko, tesigilearen argitalpenak eta tesi-proiektuak jasotako saria aipatuko ditugu. Tesiari lotutako lehen urteetako lanak hurrenkera alfabetikoari jarraituz sinatuta daude.

Jarraian, tesi-lan honi hertsiki lotutako argitalpenak zerrendatuko ditugu:

- Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. (2015) Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification. In: *Proceedings of the 7th Language & Technology Conference*. pp. 450-454. Poznań, Polonia. ISBN: 978-83-932640-7-0.

- Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. (2015) *Perpaus adberbialen agerpena, maiztasuna eta kokapena EPEC-DEP corpusean.* Barne-txostena UPV/EHU/LSI/TR 02-2015.

- Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. Salaberri, H. (2014) Simple or Complex? Assessing the Readability of Basque Texts. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. pp. 334-344, Dublin City University and Association for Computational Linguistics, Dublin (Ireland). ISBN: 978-1-941643-26-6.

- Gonzalez-Dios, I. (2014) Euskarazko testuak errazten: euskal testuen sinplifikazio automatikoa. In Aduriz, I. eta Urizar, R., eds.: *Euskal hizkuntzalaritzaren egungo zenbait ikerlerro. Hizkuntzalari euskaldunen I. topaketa*. pp. 135–149. Udako Euskal Unibertsitatea. ISBN: 978-84-8438-524-0.

- Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. (2014) Making Biographical Data in Wikipedia Readable: A Pattern-based Multilingual Approach. In: *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*. Workshop at Coling 2014, Dublin (Ireland). pp. 11–20.

- Gonzalez-Dios, I. (2014) Simplificación automática de textos en Euskera. In Ureña López, L.A., Troyano Jiménez, J.A., Ortega Rodríguez, F. J., Martínez Cámara, E. (eds.): *Actas de las V Jornadas TIMM*, Cazalla de la Sierra, España, http://ceur-ws.org. pp.45–50

- Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A. (2013). Testuen sinplifikazio automatikoa: arloaren egungo egoera. *Linguamática*. 5(2), 43–63. ISSN 1647-0818.

- Gonzalez-Dios, I. (2013) Euskarazko testuen sinplifikazio automatikoa. *Hizkuntzalari Euskaldunen I. Topaketak. Egungo ikerlerroak.*. Komunikazioa. Udako Euskal Unibertsitatea.

- Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A., Soraluze, A. (2013) Detecting Apposition for Text Simplification in Basque. *Lecture Notes in Computer Science (LNCS)* 7817, Gelbukh, A. (ed.), Computational Linguistics and Intelligent Text Processing. Springer. 13th International Conference, CICLing 2013. Part II. pp. 513–524.

- Aranzabe, M.J., Díaz de Ilarraza, A., Gonzalez-Dios, I. (2013) Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento del Lenguaje Natural*, 50, pp. 61–68. ISSN (paperezkoa): 1135-5948 ISSN (digitala): 1989-7553.

- Aranzabe, M.J., Díaz de Ilarraza, A., Gonzalez-Dios, I. (2012) First Approach to Automatic Text Simplification in Basque. In Rello, L., Saggion, H., eds.: *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, pp. 1–8, Istanbul, Turkey.

- Gonzalez-Dios, I. (2011) Euskarazko egitura sintaktikoen azterketa testuen sinplifikazio automatikorako: aposizioak, erlatibozko perpausak eta denborazko perpausak. Master tesia, Euskal Herriko Unibertsitatea (UPV-EHU).

Argitalpen horiek tesi-txosten honen kapituluekin lotu ditugu, 1.3 taulan.

Ondorengo argitalpenak, berriz, HPko beste ikerketa-lerroekin lotutakoak dira:

- Agirrezabal, M., Gonzalez-Dios, I., Lopez-Gazpio, I. (2015). Euskararen sorkuntza automatikoa: lehen urratsak. In: *I. Ikergazte Nazioarteko ikerketa euskaraz Kongresuko artikulu-bilduma* pp. 15–23. Udako Euskal Unibertsitatea. ISBN: 978-84-8438-539-4.

Kapituluak	**Argitalpenak**
2. kapitulua	Gonzalez-Dios *et al.* (2013b)
3., 4., 5. eta 6. kapituluak	Gonzalez-Dios (2011), Aranzabe *et al.* (2012a), Gonzalez-Dios (2013), Gonzalez-Dios (2014b), Gonzalez-Dios (2014a)
3., 4. eta 6. kapituluak	Aranzabe *et al.* (2013), Gonzalez-Dios *et al.* (2013a), Gonzalez-Dios *et al.* (2014a), Gonzalez-Dios *et al.* (2015b), Gonzalez-Dios *et al.* (2015a)
5. kapitulua	Gonzalez-Dios *et al.* (2014b)

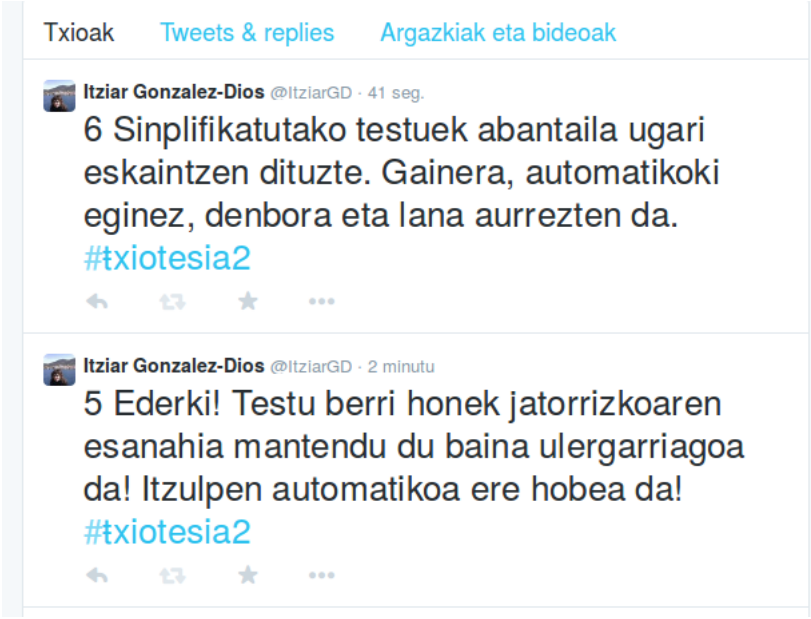
1.3 taula – Kapituluekin lotutako argitalpenak

- Aduriz I., Arriola J.M., Gonzalez-Dios I., Urizar R. (2015) *Funtzio Sintaktikoen Gold Estandarra eskuz etiketatzeko gidalerroak.* Barnetxostena UPV/EHU/LSI/TR 01-2015

- Iruskieta M., Aranzabe M. J., Diaz de Ilarraza A., Gonzalez-Dios I., Lersundi M., Lopez de Lacalle O. (2013) The RST Basque TreeBank: an online search interface to check rhetorical relations. In: *Proceedings of the 4th Workshop RST and Discourse Studies*, pp. 40-49, Sociedad Brasileira de Computaçao, Fortaleza, CE, Brasil.

- Aldabe I., Gonzalez-Dios I., Lopez-Gazpio I., Madrazo J., Maritxalar M. (2013) Two Approaches to Generate Questions in Basque. *Procesamiento del Lenguaje Natural*, 51 pp. 101-108. ISSN (paperezkoa): 1135-5948 ISSN (digitala): 1989-7553.

Tesi-lan honek Udako Euskal Unibertsitateak (UEU) antolatutako 2014ko "Txiokatu zure tesia 6 mezutan #Txiotesia2" lehiaketan "Txiolari ulergarriena" saria jaso zuen. Lehiaketa horretan tesiaren laburpena sei txiotan egin behar zen. Hona hemen sarituak izan ziren txioak:


Txioak
Tweets & replies
Argazkiak eta bideoak
Itziar Gonzalez-Dios @ItziarGD · 41 seg.
6 Sinplifikatutako testuek abantaila ugari eskaintzen dituzte. Gainera, automatikoki eginez, denbora eta lana aurrezten da. #txiotesia2
Itziar Gonzalez-Dios @ItziarGD · 2 minutu
5 Ederki! Testu berri honek jatorrizkoaren esanahia mantendu du baina ulergarriagoa da! Itzulpen automatikoa ere hobea da! #txiotesia2


Itziar Gonzalez-Dios @ItziarGD · 2 minutu
4 Testua konplexua da! Eta sinplifikatuko dut! Hitz zailak ordezkatuko ditut eta esaldiak banatuko ditut. #txiotesia2
Hvaghvavh nlknz kl jbs jh lkzj o zl mlzñ kñlz kz kmihv ljzosahfvpuiohvpa vnzjvklkc cvnb gertatzen da. Orduan, vjsahbhsalñkvalkv njkhnvjdfhap ez da ikusten. Hala ere, hcjhcajkcjlkahjvh vhbajhvbanca jbvja nvknsdbñlkmvakzn vjksasjkn osatuko da. Zhfasñjk jbiszjnçzñjvsk vlaj jvkvslkvnñlsvmzç galdetu.
View more photos and videos
Itziar Gonzalez-Dios @ItziarGD · 4 minutu
3 ErreXaili galdetuko diot, ea zer dioen! ErreXail testuak konplexuak edo sinpleak diren automatikoki esateko gai den sistema da #txiotesia2
Itziar Gonzalez-Dios @ItziarGD · 5 minutu
2 A ze esaldi luzea! Bai hitz arraroa! Ez dut ulertzn! Testua zeharo konplexua da! Sinplifikatu daiteke? #txiotesia2
Hvaghvavh nlknz kl jbs jh lkzj o zl mlzñ kñlz kz kmihv ljzosahfvpuiohvpa vnzjvklkc cvnb gertatzen denean, vjsahbhsalñkvalkv njkhnvjdfhap ikusten ez den arren, hcjhcajkcjlkahjvh vhbajhvbanca jbvja nvknsdbñlkmvakzn vjksasjkn osatuko da. Zhfasñjk jbiszjnçzñjvsk vlaj jvkvslkvnñlsvmzç itaundu.
View more photos and videos
Itziar Gonzalez-Dios @ItziarGD · 9 minutu
1 Egitura sintaktiko konplexuen identifikazioa eta sinplifikazio automatikoa euskararen tratamendu automatikoan. #txiotesia2

2

Aurrekariak: testuen konplexutasunaren analisia eta testuen sinplifikazio automatikoa

Kapitulu honetan tesi-lan honen aurrekariak aurkeztuko ditugu. Horretarako, testuen sinplifikazioa zer den eta irakaskuntzan nola ulertu den azaldu ostean, Hizkuntzaren Prozesamenduan egin diren lanetan murgilduko gara; alegia, testuen sinplifikazio automatikoa izango dugu aztergai. TSAri zuzendutako testuen konplexutasunaren analisia ere zer den azalduko dugu.

2.1 Sarrera

Testuen sinplifikazioak testu bat sinpleagoa lortzea du helburu jatorrizko testuaren esanahiari eutsiz; egitura eta hitz konplexuak ordezkatuz sortzen den testua irakurterrazagoa izan behar da irakurle jakin batzuentzat. Aipatu dugu testu sinpleek abantaila ugari eskaintzen dizkietela, bai pertsonei, bai HPko tresnei (Chandrasekar *et al.*, 1996; Siddharthan, 2002; Aluísio eta Gasperin, 2010).

HPari dagokionez, ikerketa-lerro hau azken urteetan garrantzitsua bihurtu da, ingeleserako ez ezik beste hainbat hizkuntzatarako ere proposatu direlako sistemak, eta metodo eta teknika berri ugari argitaratzen ari direlako. Testuak eskuz sinplifikatzeak edo egokitzeak lan handia eta garestia eskatzen du; hala, HPko tresnak erabilita testuak sinplifikatzean lana erraztu eta azkartu egiten da.

Testuen eskuzko sinplifikazioa irakaskuntzan eta batez ere atzerriko hizkuntzen didaktikan landu izan da. Irakaskuntzan testu sinplifikatuak erabiltzearen helburua ulermena areagotzea eta karga kognitiboa leuntzea da (Crossley *et al.*, 2012).

Atzerriko hizkuntzen irakaskuntzan, hain zuzen ere, Allen-ek (2009) eta Crossley *et al.*ek (2012) sinplifikatzeko bi aukera daudela azaltzen dute:

- Egituraren araberako sinplifikatzea: sinplifikazio-mota hori mailakatutako irakurketan erabiltzen da eta maila jakin bakoitzari dagozkion hitz zerrendak eta egitura sintaktikoen zerrendak erabilita gauzatzen da. Kasu batzuetan, testuen konplexutasun-maila neurtzen duten *readability* formuletan oinarritzen da. Formula horiek esaldien eta hitzen luzeran oinarritzen diren algoritmoak dira eta konplexutasunaren analisian baliagarriak badira ere, atzerriko hizkuntzen didaktikan formula horietan oinarritzea kritikatu izan da. Formula horiek 2.2 atalean azalduko ditugu zehatzago. Ikaslearen hizkuntza-ezagutza pixkanaka handitzen doan heinean eta mailan aurrera egin ahala, egitura eta hitz berriak sartzea da sinplifikazio-mota horren helburua.

- Intuitiboki sinplifikatzea: maila jakin bakoitzari dagozkion hitz-zerrendak eta egitura sintaktikoen zerrendak kontuan izan ditzaketen arren, oro har, intuizioari jarraiki sinplifikatzen da. Hau da, testua sinplifikatzen ari den pertsona bere irakasle- eta hizkuntza ikasle-esperientzian oinarritzen da testuak sinplifikatzeko.

Young-ek (1999), berriz, sinplifikatzeko hainbat metodo aipatzen ditu:

- Linguistikoki sinplifikatzea: testua berridaztea esaldiak laburragoak egiteko; esaera idiomatikoak ezabatzea edo parafraseatzea; hitz espezializatuak eta maiztasun gutxikoak ekiditea, eta sintaxi konplexua berrikustea perpaus bakunak sortzeko.

- Materia sinplifikatzea: testua laburtzea paragrafoak edo atalak kenduz.

- Glosen bitartez sinplifikatzea: itzulpenak edo definizioak gehitzea.

- Prozesamendu kognitiboetan oinarritutako aldaketak eta elaborazioak eginez sinplifikatzea.

Simensen-ek (1987) atzerriko hizkuntzetako liburuak egokitzeko argitaletxeek dituzten **gidalerroak** aztertu zituen, eta gidalerro horietan oinarrituta, egokitzapenerako hiru printzipio (kontrol-printzipioak) aurkezten ditu: informazioaren, hizkuntzaren eta diskurtsoaren kontrola. Hizkuntzaren kontrolaren atalean, zerrendetan (egituraren araberako sinplifikazioa) eta intuizioan (intuitiboki sinplifikatzea) oinarritutako hurbilpenak aurkezten ditu, Crossley *et al.*ek (2012) ere beranduago azaldu bezala.

Atzerriko hizkuntzen didaktikaren arloan, testu sinplifikatuak edo originalak erabiltzeak dituen abantailak ere aztertu izan dira, nahiz eta emaitzak kontrajarriak diren. Esaterako, Young-en (1999) azterketaren arabera, testu sinplifikatuak ez dira baliagarriak ikasleak irakurketa globala egiten ari badira; are gehiago, kalterako izan daitezke. Oh-ren (2001) arabera, ordea, irakurmenaren ulermen globala erraztu egiten dute. Crossley *et al.*ek (2012), berriz, uste dute testu sinplifikatuek erreferentziakidetasun bidezko kohesio altuagoa eta lokailu eta hitz ezagun gehiago eskaintzen dizkietela ingelesa ikasten ari diren ikasleei. Dena dela, arlo honetako egile gehienek gaian sakondu behar dela adierazten dute (Young, 1999; Allen, 2009; Crossley *et al.*, 2012).

HPan hartu den sinplifikatzeko ildoa linguistikoki edo estrukturalki sinplifikatzearena izan da; zehazki, bi sinplifikazio-mota landu dira: sintaktikoa eta lexikala. Testuen sinplifikazio automatikoa Siddharthan-ek (2002) berridazketa-prozesu bat bezala definitzen du konplexutasun lexikala eta sintaktikoa gutxitzeko. Aluísio eta Gasperin-ek (2010), berriz, HPko ikerketalerro bezala definitzen dute PorSimples[1] proiektuan. Bertan Brasilgo portugesezko testuen ulermena areagotzeko, fenomeno lexikalak eta sintaktikoak sinplifikatzen dituzte; lehendabizi, pertsona gutxik ulertzen dituzten hitzak ezagunagoekin ordezkatzen dituzte eta bigarrenik, esaldiak banatu eta esaldien egitura aldatzen dute.

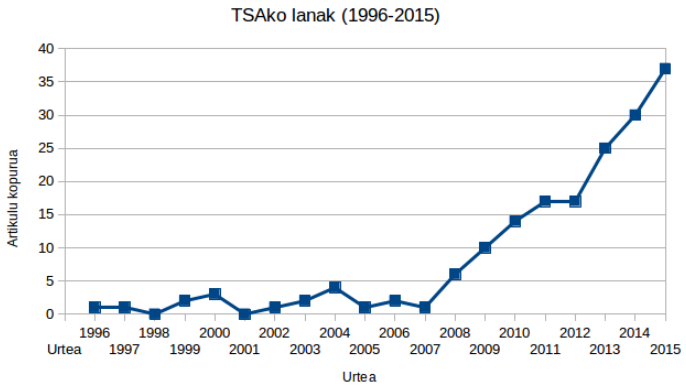
HPko testuen sinplifikazioan, hasierako lanak ingeleserako egin ziren. Lehen lana Chandrasekar *et al.*ena (1996) izan zen eta bertan TSArako motibazioak azaldu zituzten. Hasierako beste lanen artean, *Practical Simplification of English Text* (PSET) proiektuan[2] (Carroll *et al.* 1998) egindakoa aurki daiteke. Proiektu horretan hizkuntzarekin arazoak zituztenei eta, batez ere, afasia zuten pertsonei zuzendutako sinplifikazioa egin zuten Carroll *et al.*ek

[1] http://caravelas.icmc.usp.br/wiki/index.php/Principal (2011ko irailean atzituta)

[2] http://www.informatics.sussex.ac.uk/research/groups/nlp/projects/pset.php (2013ko maiatzean atzituta)

(1999). Siddharthan-ek (2002), berriz, testuen sinplifikazio automatikorako oinarrizko arkitektura finkatu zuen.

Testuen sinplifikazio automatikoa oso garrantzitsua bilakatu da 2009tik aurrera eta beste hizkuntzetara zabaltzeaz gain, metodo eta teknika berri ugari argitaratu dira, batez ere metodo estatistikoetan eta ikasketa automatikoan oinarrituz. Azken hogei urteetan izan den lanen hazkundearen grafikoa ikus dezakegu 2.1 irudian. Testuen sinplifikazioan egin diren lanak Shardlow-en (2014) eta Siddharthan-en (2014) lanetan ere bildu dira. Aipatu nahi dugu azken urteotan hainbat tesi egin direla (Siddharthan, 2003; Petersen, 2007; Temnikova, 2012; Rello, 2014; Štajner, 2015; Brunato, 2015) ikerketa-lerro horretan.



2.1 irudia – TSAko lanen kopurua

Esan bezala, TSArako sistema gehienak ingeleserako proposatu eta egin dira; horien artean ditugu Siddharthan-ena (2006) eta Zhu *et al.*ena (2010). Azken urteotan beste hizkuntzetarako ere egin dira: japoniera (Inui *et al.*, 2003; Kajiwara eta Yamamoto, 2015), Brasilgo portugesa (Candido *et al.*, 2009; Gasperin *et al.*, 2009a; Aluísio eta Gasperin, 2010), suediera (Rybing *et al.*, 2010; Keskisärkkä, 2012; Rennes eta Jönsson, 2015), arabiera (Al-Subaihin eta Al-Khalifa, 2011), gaztelania (Saggion *et al.*, 2011, 2013), frantsesa (Seretan, 2012; Brouwers *et al.*, 2012), italiera (Barlacchi eta Tonelli, 2013), daniera (Klerke eta Søgaard, 2013), bulgariera (Lozanova *et al.*, 2013), koreera (Chung *et al.*, 2013), hindiera (Mishra *et al.*, 2014) eta alemaniera (Suter, 2015).

Chandrasekar *et al.*en (1996) lanean, TSA gizakientzat eta HPko tresnentzat erabilgarria eta onuragarria dela esaten da. Urteak pasa ahala, lan horretan proposatutako aplikazioak (analizatzaile sintaktikoentzat, itzulpen automatikorako, informazio-berreskupenerako, laburpen automatikoak egiteko eta testuaren argitasuna bermatzeko) gauzatu dira. Jarraian ikusiko ditugu gizakiei eta tresnei zuzenduta egin diren TSAko lanak:

- **Gizakiei zuzendutakoak.** Testu sinpleek informazioa irisgarriagoa bihurtzen dute eta horrela testuak ulertzea errazagoa da. Jarraian, gizakientzat egin diren lanak zerrendatuko ditugu:
  - Urritasunak dituztenenak (Carroll *et al.*, 1999): afasikoak (Carroll *et al.*, 1998; Max, 2005; Devlin eta Unthank, 2006), jaiotzetiko entzumen-arazoak dituztenak (Inui *et al.*, 2003; Vettori eta Mich, 2011; Lozanova *et al.*, 2013; Chung *et al.*, 2013), irakurtzeko arazoak dituztenak (Bautista *et al.*, 2012b), dislexikoak (Rello *et al.*, 2013a), adimen-urritasunak dituztenak (Saggion *et al.*, 2011; Bott eta Saggion, 2012; Saggion *et al.*, 2015a; Fajardo *et al.*, 2013), eta autistak (Evans *et al.*, 2014; Barbu *et al.*, 2015)
  - Atzerriko hizkuntzen ikasleak (Petersen, 2007; Burstein, 2009)
  - Haurrak (De Belder eta Moens, 2010; Brouwers *et al.*, 2012; Barlacchi eta Tonelli, 2013)
  - Gutxi alfabetatuak (Candido *et al.*, 2009; Al-Subaihin eta Al-Khalifa, 2011)
- **Oinarrizko tresnei edo HPko aplikazio aurreratuei zuzendutakoak.** Esaldi luzeak eta konplexuak dituzten testuak zailagoak izaten dira automatikoki prozesatzeko; esaldi laburragoak eta sinpleak erabiltzen diren kasuetan, aldiz, tresnen eta aplikazio aurreratuen errendimendua hobea da. Hori dela-eta, testuen sinplifikazioa aurreprozesu bezala erabil daiteke performantzia igotzeko. Hobekuntza hori oinarrizko tresnetan eta aplikazio aurreratuetan gertatzen da. TSA edo esaldien sinplifikazioa ikerketa-lerro edo sistema hauetan erabili da:
  - Oinarrizko tresnak: dependentzia-gramatikan oinarritutako analizatzaile sintaktiko edo *parser*ak (Chandrasekar *et al.*, 1996); rol semantikoen etiketatzailea (Vickrey eta Koller, 2008); arlo berezituetako analizatzailea, adibidez biomedikuntzako testuetakoa

(Jonnalagadda *et al.*, 2009); eta ahozko hizkuntza ulertzen duten sistemak (Tur *et al.*, 2011)

- Informazioa bilatzeko eta erauzteko sistemak: informazioa bilatzeko aplikazioak (Beigman Klebanov *et al.*, 2004), informazio-erauzketa (Jonnalagadda eta Gonzalez, 2010b; Evans, 2011), gertaeren erauzketa (Buyko *et al.*, 2011) eta erlazioen erauzketa (Minard *et al.*, 2012)
- Laburpen automatikoak egiteko sistemak (Lal eta Rüger, 2002; Siddharthan *et al.*, 2004; Blake *et al.*, 2007; Vanderwende *et al.*, 2007; Bawakid eta Oussalah, 2011; Silveira Botelho eta Branco, 2012; Finegan-Dollak eta Radev, 2015)
- Azpitituluak egiteko sistemak (Daelemans *et al.*, 2004)
- Itzulpen automatikoa (Doi eta Sumita, 2004; Poornima *et al.*, 2011; Tyagi *et al.*, 2015)
- Galdera-sorkuntza sistemak (Heilman eta Smith, 2010; Bernhard *et al.*, 2012; Al Tarouti *et al.*, 2015)
- Corpus paraleloetan hitzak lerratzeko sistemak (Srivastava eta Sanyal, 2012)

Testuak helburu-talde ezberdinetara bidera daitezkeen bezala, erabiltzaile horien arabera ere maila ezberdinetan sinplifika daitezke testuak. Adibidez, Brasilgo portugesez sinplifikazio-maila[3] naturala (*natural*) eta bortitza (*strong*) proposatu dira (Gasperin *et al.*, 2009a). Sinplifikazio-maila horiek jendearen alfabetatze-mailaren araberakoak dira eta 4. kapituluan izango ditugu aztergai. Gaztelaniaz ere sinplifikazio gogorrak (*heavy*) eta leunak (*soft*) biltzen dituzten corpusak osatu dira (Štajner *et al.*, 2015b). Sinplifikazio-maila horien arteko diferentzia testuak sinplifikatzean erabilitako gidalerroetan datza.

Fenomeno jakinak tratatzen dituzten lanak ere badira: entitate konposatuak biomedikuntzako testuetan (Wei *et al.*, 2014), zenbakizko adierazpenak (Bautista *et al.*, 2015), perpaus koordinatuak (Evans, 2011) eta erlatibozko perpausak (Saini *et al.*, 2015).

[3]Euskarazko terminoarekin batera egileek ingelesez darabiltena ematea erabaki dugu sinplifikazio-mailak adierazteko.

Domeinuaren araberako testuak lantzen dituzten lanak ere aurki daitezke: atzerriko hizkuntzen irakaskuntza eta testuliburuen pertsonalizazioa (Petersen, 2007; Burstein, 2009; Nunes *et al.*, 2013), medikuntza-testuak (Damay *et al.*, 2006; Jonnalagadda eta Gonzalez, 2010a; Kandula *et al.*, 2010; Peng *et al.*, 2012), larrialdien kudeaketa (Temnikova *et al.*, 2012), webeko dokumentuak (Chung *et al.*, 2013), administrazio-testuak [patenteak (Sheremetyeva, 2014) eta parlamentuko dokumentuak (Collantes *et al.*, 2015)], zientzia-kazetaritza (Kim *et al.*, 2015) eta testu historikoak (Vertan eta von Hahn, 2014).

Sarrerarekin amaitzeko, testuen sinplifikazio automatikoa egiten duten produktuak aipatuko ditugu. Horietako bat SIMPLIFICA sistemaren (Scarton *et al.*, 2010) aplikazioa da, zeinak testu bat emanda konplexutasuna kalkulatu eta sinplifikatzen duen. Sinplifikatzean, tresnak proposatutako sinplifikazioa emateaz gain, editatzeko aukera ematen du, testuak sinplifikatzen dituzten egileei lana erraztuz. Testu-editoreetan sinplifikazio lexikala eta sintaktikoa ere integratu dira (Hervás *et al.*, 2014). Irakaskuntzan erabiltzeko badira Adaptive eBook (Dingli eta Cachia, 2014) sistema, zientzia-kazetaritzarako DeScipher (Kim *et al.*, 2015) editorea, dislexikoei testua eskuragarriago egiteko DysWebxia (Rello *et al.*, 2013d) aplikazioa eta afasia dutenei laguntzeko Open Book (Barbu *et al.*, 2015) tresna[4]. Sistema baten demostrazioa ere badago webean (Ferrés *et al.*, 2015).

2.2 TSA helburu duen testuen konplexutasunaren analisia

Testuen konplexutasunaren analisia (TKA) HPko ikerketa-lerro bat da. Testu bat emanda, testu horren zailtasun-maila adieraztea du helburu. Testuen konplexutasuna jakitea gai garrantzitsua da iraskuntzan XX. mendean, eta HPan azkeneko urteetan lerro arrakastatsua bihurtu da.

Konplexutasuna analizatzeko formula edo metrika klasikoek [besteak beste, Flesh formulak (Flesch, 1948), Dale-Chall formulak (Chall eta Dale, 1995) eta Gunning FOG indizeak (Gunning, 1968], azaleko ezaugarriak (hitz- eta silaba-kopuruak), ezaugarri lexikalak eta maiztasunak hartzen dituzte kontuan, eta hizkuntzekiko lotura estua dute. HPko teknikek, bestalde, ezaugarri

[4]Tresna hori FIRST proiektuan sortu da eta ingeleserako, gaztelaniarako eta bulgarierako da baliagarria.

gehiago eta konplexuagoak har ditzakete kontuan. Gainera, HPko lanetan (Si eta Callan, 2001; Petersen eta Ostendorf, 2009; Feng, 2009) erakutsi da formula klasikoak ez direla fidagarriak.

TKA testuen sinplifikazio automatikoan aurreprozesu edo ebaluazio bezala erabili da, adibidez ingelesez (Feng *et al.*, 2010; Vajjala eta Meurers, 2014b), portugesez (Aluísio *et al.*, 2010), italieraz (Dell'Orletta *et al.*, 2011), alemanez (Hancke *et al.*, 2012) eta gaztelaniaz (Štajner eta Saggion, 2013; Štajner *et al.*, 2015c). Aurreprozesuaren kasuan, sistema horien helburua da testu bat sinplea edo konplexua den erabakitzea; konplexua izanez gero, testua sinplifikatu egin beharko da. Ebaluazioaren kasuan, sinplifikatutako testua sinplea den egiaztatzen da.

Testuen konplexutasuna analizatzeko erabili diren metodoei dagokienez, ingelesez, Si eta Callan-ek (2001) eredu estatistikoak eta ezaugarri klasikoak konbinatu dituzte; zehazki, unigrametan oinarritutako hizkuntza-ereduak eta esaldiaren luzera eta hitz bakoitzeko silaba-kopuruak erabili dituzte. Graesser *et al.*ek (2004) garatutako Coh-Metrix tresna, berriz, ezaugarri anitzetan eta diskurtsoaren mailetan (narratibotasuna, hitzaren zehaztasuna edo izenen gainjartzea) oinarritzen da. 3.0 bertsioan[5] 108 indize eskuragarri daude. Pitler eta Nenkova-k (2008) ezaugarri lexikalak, sintaktikoak eta diskurtsiboak erabili dituzte; gainera, ezaugarri diskurtsiboen garrantzia azpimarratu dute. Schwarm eta Ostendorf-ek (2005) hizkuntza-eredu estatistikoen ezaugarriak, analisiaren ezaugarriak eta ezaugarri tradizionalak konbinatu dituzte euskarri bektoredun makinak erabiliz.

Beste hizkuntzetako sistemetan, hizkuntzarekiko espezifikoak diren ezaugarriak hartu dituzte kontuan: txineraz, trazu-kopurua (Pang, 2006); japonieraz, karaktere ezberdinak (Sato *et al.*, 2008); alemanez, hitz-sorkuntza (vor der Brück *et al.*, 2008); frantsesez, *passé simple* aldia (François eta Fairon, 2012) eta ortografia-antzekotasuna (Gala *et al.*, 2013); eta suedieraz, baliabide lexikalak (Sjöholm, 2012; Falkenjack *et al.*, 2013). Brasilgo portugesez Coh-Metrix egokitu dute Scarton eta Aluísio-k (2010) eta arabieraz, berariaz prestatuko formulak erabili dituzte (Al-Ajlan *et al.*, 2008; Daud *et al.*, 2013). Hurrenkera librea, buru azkenak eta morfologia aberatsa duten hindi eta bangla hizkuntzetarako, bi formula proposatu dituzte Sinha *et al.*ek (2012) ingeleserako formuletan oinarrituta. Beste sistema batzuetan ikasketa automatikoko teknikak soilik erabili dituzte, adibidez txineraz (Chen *et al.*,

[5] http://cohmetrix.memphis.edu/cohmetrixpr/cohmetrix3.html (2014ko urtarrilean atzituta)

2011).

Testuen sinplifikazio automatikoan aurreprozesu bezala erabiltzen diren sistema horiek ezaugarri linguistikoen analisia egiten dute eta horietan oinarrituta, testuak sinpleak edo konplexuak diren erabakiko duten sailkatzaileak eraikitzeko, ikasketa automatikoa erabiltzen dute. Sistema horiek ingeleserako (Feng *et al.*, 2010), portugeserako (Aluísio *et al.*, 2010), italierarako (Dell'Orletta *et al.*, 2011) eta alemanerako (Hancke *et al.*, 2012) sortu dituzte. Sistema horiek guztiak 5. kapituluan izango ditugu hizpide.

Badira dokumentu mailako konplexutasuna aztertzeaz gain, esaldi mailakoa ere aztertzen duten sistemak (Dell'Orletta *et al.*, 2011; Vajjala eta Meurers, 2014a, b). Beste sistema batzuek hitzen konplexutasuna (Wilkens *et al.*, 2014) aztertzen dute.

TKA domeinu ezberdinetara ere bidera daiteke: administrazio-testuak, medikuntza-testuak, irakaskuntza-testuak eta abar. Esaterako, medikuntza-testuak sinplifikatzeko, ezaugarri konplexuak ikasketa automatikoaren bitartez aztertzen dituzte Kauchak *et al.*ek (2014).

TKAri buruzko gako interesgarriak DuBay-ren (2004) lanean aurki daitezke. Metodoen analisia eta sistemen azterketa, berriz, Benjamin-en (2012) eta Zamanian eta Heydari-n (2012) ikus daiteke.

2.3 TSArako eta TKArako baliabideak: corpusak eta datu-multzoak

TSAren atazan, HPko beste atazetan bezala, corpusak oso baliabide garrantzitsuak dira. Bi motakoak dira bereziki erabilienak: corpus paraleloak eta corpus ez-paraleloak.

Corpus paraleloetan jatorrizko testua eta testu sinplea lerratuta daude, esaldiz esaldi; hau da, jatorrizko testuko esaldi bakoitzak bere baliokide sinplea du. Corpus horien helburua, oro har, jatorrizko testu batetik testu sinplifikatua lortzeko egin diren eragiketak aztertzea da eta horietan zein fenomeno gertatzen diren azaltzeko etiketatze-eskemak osatu dira (Bott eta Saggion, 2014; Brunato *et al.*, 2015). Horrela etiketatutako corpus paraleloen ezaugarriak 7. kapituluan ikusiko ditugu.

Corpus paraleloak hainbat hizkuntzatarako sortu dira, hala nola, ingeleserako (Petersen eta Ostendorf, 2007; Pellow eta Eskenazi, 2014; Xu *et al.*, 2015), Brasilgo portugeserako (Caseli *et al.*, 2009), gaztelaniarako (Bott eta

Saggion, 2011; Štajner *et al.*, 2013; Bott eta Saggion, 2014; Štajner, 2015), danierarako (Klerke eta Søgaard, 2012) eta italierarako (Brunato *et al.*, 2015). Corpus paraleloak adituek eskuz sortu izan dituzte erreferentziazko corpusa hartuta eta erregela jakin batzuk aplikatuta (Collados, 2013) edo erregela-multzo ezberdinak konparatuta (Mitkov eta Štajner, 2014). Corpus batzuetan, gainera, sinplifikazio-mailaren edo hurbilpen ezberdinaren arabera egindako testuak aurki daitezke (Caseli *et al.*, 2009; Brunato *et al.*, 2015; Štajner, 2015; Xu *et al.*, 2015).

Corpus ez-paraleloek, berriz, lerratu gabe gordetzen dituzte testu sinpleak eta jatorrizkoak. Corpus horiek batez ere testu sinpleen eta testu konplexuen ezaugarriak ezagutzeko erabili ohi dira, eta, batez ere, TKAn erabili dira. Modu horretan, Dell'Orletta *et al.*ek (2011) italierarako eta Hancke *et al.*ek (2012) alemanerako, haurrei zuzendutako egunkarietako eta aldizkarietako testuak testu sinple bezala jaso dituzte batetik, eta prentsa arrunteko testuak bestetik. Alemanerako ere corpus ez-paraleloa eraiki dute Klaper *et al.*ek (2013) webeko testuak erabiliz. Medikuntza arloan, eta zehatzago esanda, erradiologiako txostenetan oinarrituta, Kvistab eta Velupillaia-k (2013) suedierako corpusa osatu eta aztertu dute.

Wikipedia entziklopedia ere corpus moduan erabili da. *Wikipedia*ren ingelesezko jatorrizko bertsioaz[6] gain, *Simple English*[7] edo ingeles errazean idatzitako bertsioa dago eskuragarri (Yatskar *et al.*, 2010; Woodsend eta Lapata, 2011b; Coster eta Kauchak, 2011a; Shardlow, 2013b). Frantseserako, Brouwers *et al.*ek (2012) frantsesezko *Wikipedia* erabili dute jatorrizko testuak lortzeko eta *Vikidia*[8], 8-13 urte tartekoei zuzendutako entziklopedia, testu sinpleak lortzeko. Bigarren corpus batean, jatorrizko ipuinak eta frantsesa ikasten ari direnentzat egokitutako bertsioak jaso dituzte Brouwers *et al.*ek (2014).

Wikipedia baliatu da datu-multzoak definitzeko eta horrekin sinplifikazioa egiten duten tresnak eta sistemak garatzeko eta ebaluatzeko. Zhu *et al.*ek (2010), esaterako, *Wikipedia*tik eta *Simple Wikipedia*tik hartutako 65.133 artikuluko datu-multzoa bildu dute. Artikuluak parekatzeko *language link* edo hizkuntzen arteko esteka jarraituz, Wikimediako *dump files* erabili dute. Halaber, *Wikipedia* eta *Simple Wikipedia*n oinarrituta, jatorrizko eta *Simple Wikipedia*ko 137.000 esaldi lerratu dituzte Coster eta Kauchak-ek (2011b).

[6] http://en.wikipedia.org/wiki/Main_Page (2013ko irailean atzituta)
[7] http://simple.wikipedia.org/wiki/Main_Page (2013ko irailean atzituta)
[8] http://fr.vikidia.org/wiki/Accueil (2013ko irailean atzituta)

2007ko SemEval-eko lexikoaren ordezkapena (*lexical substitution*) atazarako sortu zen datu-multzoa oinarri hartuta, etiketatze-prozesua eta etiketatzaileen arteko adostasun-emaitzak azaldu dituzte De Belder eta Moens-ek (2012).

Corpusak (paraleloak zein ez-paraleloak) lerrokatzeko algoritmoak ere (Bott eta Saggion, 2011; Klaper *et al.*, 2013; Hwang *et al.*, 2015) aurkeztu dira, eta corpusen analisiak ere egin dira (Štajner *et al.*, 2014a).

2.4 TSArako sistemak

HPan egiten diren testuen sinplifikazioak bi mota nagusikoak dira: sinplifikazio sintaktikoa eta sinplifikazio lexikala. Bi horiek TSAko azpiatazatzat hartu izan ohi dira.

Sinplifikazio sintaktikoak testu baten konplexutasun gramatikala murriztea du helburu. Horretarako, egitura sintaktiko konplexuak sinpleagoez ordezkatzen dira (Siddharthan, 2006). Sinplifikazio sintaktikoari eman zitzaion garrantzirik handiena hasiera batean testuen sinplifikazio automatikoan. Hasierako lan gehienak eskuzko erregeletan oinarritutakoak (ezagutza linguistikoan) ziren, baina azken urteetan *Wikipedia* bezalako baliabideei esker, metodo estatistikoak ugaritu egin dira. Azken urte hauetan, sinplifikazio lexikalak ere bere tokia hartu du. Sinplifikazio-mota horren helburua hitzen ulergarritasuna areagotzea da, hitz konplexuak edo maiztasun gutxikoak baliokide diren hitz ezagunagoekin, sinonimoekin edo sintagmekin ordezkatuz (Specia *et al.*, 2012). Badaude sinplifikazio sintaktikoa eta lexikala uztartzen dituzten sistemak ere. Sistema horiek bi motako sinplifikazioak egiten dituzte, hau da, testuaren konplexutasun lexikala eta sintaktikoa murrizten dute.

Testuak automatikoki sinplifikatzeko metodoei eta teknikei dagokienez, HPko beste atazetan bezala, hiru multzo nagusi bereizi behar dira: ezagutza linguistikoan oinarritutakoak, estatistikan eta ikasketa automatikoan (datuetan) oinarritutakoak eta aurreko biak batzen dituzten sistema hibridoak. Azken urteotan estatistikan oinarritutako metodoek lekua irabazi diete ezagutza linguistikoan oinarritutako sistemei, eskuzko erregeletan oinarritutako sistemak eraikitzeak denbora asko eskatzen baitu.

Hiru multzo horietaz gain, badira TSA itzulpen prozesu bezala ulertzen dutenak; hau da, jatorrizko testuak testu sinplifikatu bihurtzea itzulpena balitz bezala ulertzen da, estatistikoa nahiz eskuzko erregeletan oinarritutakoa. Horrela, jatorrizko testuaren hizkuntza *A iturri-hizkuntza*ren pareko izango

litzateke eta, testu sinplearen hizkuntza *B helburu-hizkuntza*rena. Sistema horiek batez ere teknika estatistikoetan oinarritzen dira.

Sistemen arkitekturei dagokienez, sistema gehienek modulu hauek dituzte: i) analizatzailea: bertan analisia egiten da, sinplifikazioaren aurreprozesua dena; izan ere, testua analizatu gabe ezin da testua sinplifikatu. Analisi hori gauzatzeko analizatzaile sintaktikoak osagai-ereduan edo dependentzia-ereduan oinarritzen dira. ii) Sinplifikatzailea edo transformatzailea: testuak sinplifikatzeaz arduratzen den modulua da, bere gain metodo eta teknikak hartzen dituena. iii) Testuen kohesioa bermatzen duen modulua. Analisia dela-eta, lan gehienetan testuak sintaktikoki analizatzen badira ere, lan berrietan hasi dira analisi semantikoaren gainean lan egiten.

Hurrengo azpiataletan, batetik, sinplifikazio-motak azalduko ditugu, bestetik TSAko sistemen arkitekturen eboluzioa ikusiko dugu, eta azkenik, sistemak sailkatuko ditugu erabiltzen duten teknikaren arabera.

2.4.1 Sinplifikazio-motak

Esana dugu bi sinplifikazio-mota nagusi daudela: sinplifikazio sintaktikoa eta sinplifikazio lexikala. Horietaz gain, bestelako sinplifikazioak eta domeinuetara egokitutako sinplifikazioak ere badira.

Sinplifikazio sintaktikoa

Sinplifikazio sintaktikoak, aurretik aipatu bezala, konplexutasun gramatikala murriztea du helburu, jatorrizko testuaren informazioari eta esanahiari eutsiz (Siddharthan, 2006). Ataza hori egiteko, hizkuntzalariek edo adituek egindako erregelak, corpusetatik ikasitako transformazioak eta irakurketa errazeko gomendioak erabili dira, besteak beste. Sinplifikazio sintaktiko automatikoaren adibideak 2.1 taulan emango ditugu ingelesez, Brasilgo portugesez, gaztelaniaz, frantsesez eta italieraz.

Hizkuntza	**Jatorrizko esaldia**	**Sinplifikatutako esaldiak**
Ingelesa Siddharthan (2002)	*Needing* money to pay my rent, I forced myself to beg my parents.	I needed money to pay my rent. I forced myself to beg my parents.

(Jarraipena hurrengo orrialdean)

Hizkuntza	Jatorrizko esaldia	Sinplifikatutako esaldiak
Brasilgo portugesa Specia *et al.* (2008)	Vários produtos brasileiros enfrentam barreiras para entrarem nos EUA, *ao passo que* o mercado brasileiro está basicamente aberto.	Vários produtos brasileiros enfrentam barreiras para entrarem nos EUA. Mas o mercado brasileiro está basicamente aberto.
Gaztelania Bott eta Saggion (2012)	Los participantes (...) recibirán como obsequio un libro *editado* por el Ayuntamiento (...).	Los participantes (...) recibirán como obsequio un libro. Este libro está editado por el Ayuntamiento (...).
Frantsesa Seretan (2012)	Il faut favoriser l'éducation des enfants et des adultes pour une prise de conscience des risques, *mais* aussi développer la sécurisation des réseaux routiers (...).	Il faut favoriser l'éducation des enfants et des adultes pour une prise de conscience des risques. Mais il faut aussi développer la sécurisation des réseaux routiers (...).
Italiera Barlacchi eta Tonelli (2013)	Ernesta *stava mangiando* la torta *con i suoi amici.*	Ernesta mangia la torta.

2.1 taula – Sinplifikazio sintaktikoaren adibideak hainbat hizkuntzatan

Sinplifikazio sintaktikoa egitean, hainbat fenomeno linguistiko daude jomugan. 2.2 taulan hainbat lanetan landutako fenomenoak jaso ditugu, egileen jatorrizko terminologiari eutsiz. Oro har, laburbilduz, lantzen diren fenomenoak perpaus koordinatuak, mendeko perpausak, aposizioak eta ahots pasiboa dira.

Lana	Fenomenoak
Ingelesa	
Siddharthan (2002)	Perpaus erlatiboak, perpaus adberbialak (menderagailurik gabeak), perpaus koordinatuak, mendeko perpausak (menderagailudunak), perpaus korrelatuak, partizipio-sintagmak, aposizioak eta ahots pasiboa
De Belder eta Moens (2010)	Aposizioak, erlatibozko perpausak, *prefix subordination*[9] eta *infix coordination and subordination*[10]
Siddharthan (2010)	Konektoreak

(Jarraipena hurrengo orrialdean)

[9]Fenomeno horrekin adierazten dute zein diren sinplifikatzean txertatze-elementua gehitu behar zaien mendeko perpausak.

[10]Fenomeno horrekin adierazten dute zein diren sinplifikatzean banaketa soilik egiten diren mendeko perpausak eta koordinatuak.

Lana	Fenomenoak
Evans (2011)	Perpaus koordinatuak
Poornima *et al.* (2011)	Perpaus koordinatuak, mendeko perpausak eta erlatibozko izenordaina duten perpausak
Siddharthan (2011)	Koordinazioa, mendeko perpausak, erlatibozko perpausak, aposizioak eta ahots pasiboa
Peng *et al.* (2012)	Koordinazioa, erlatibozko perpausak eta aposizioak
Brasilgo portugesa	
Aluísio *et al.* (2008b)	Aposizioak, erlatibozko perpausak, perpaus koordinatuak, mendeko perpausak eta ahots pasiboa
Arabiera	
Al-Subaihin eta Al-Khalifa (2011)	Elipsia, banatutako subjektuak, objektuak eta aditzak, eta esaldi pasiboak
Gaztelania	
Bott *et al.* (2012b)	Erlatibozko perpausak, gerundiozko eta partizipiozko egiturak, perpaus koordinatuak eta objektuen koordinazioa
Frantsesa	
Seretan (2012)	Unibertso-sartzaileak, aposizioak, perpaus osagarriak, perpaus koordinatuak, mendeko perpausak, objektu funtzioa duten erlatibozko perpausak, gerundiozko perpausak, adjuntu luzeak, izenen ondorengo modifikatzaile luzeak, partizipiozko modifikatzaile luzeak, esaldi oso laburrak eta enfatizazioa

2.2 taula – Landutako fenomeno sintaktikoak

Fenomeno horiek sinplifikatzeko, eragiketak egin behar dira. Sinplifikazio-eragiketa horiek 2.3 taulan jaso ditugu. Eragiketak izendatzean, egileek erabili dituzten terminoak eman ditugu.

Lana	Eragiketak
Japoniera	
Inui *et al.* (2003)	Parafrasi sintaktikoak
Ingelesa	
Daelemans *et al.* (2004)	Sintagmak ezabatu; testuko hitzak kopiatu, ezabatu edo ordezkatu
Zhu *et al.* (2010)	Perpausak banatu, hitzak ezabatu, ordenatu eta sintagma/hitz ordezkatu
Bawakid eta Oussalah (2011)	Esaldiak banatu eta esaldiak trinkotu

(Jarraipena hurrengo orrialdean)

Lana	Eragiketak
Vu *et al.* (2014)	Banaketa, ezabaketa, hurrenkera-aldaketa eta ordezkapena
Brasilgo portugesa	
Aluísio *et al.* (2008b)	Esaldiak banatu, diskurtso-markatzailea aldatu, ahots pasiboa aktibo bihurtu, perpausen hurrenkera aldatu, hurrenkera kanonikoa errespetatu, adizlagunak galdegai/ez-galdegai bihurtu eta ez sinplifikatu
Gaztelania	
Saggion *et al.* (2011)	Manipulazio sintaktikoak: banaketa, ezabaketa, txertaketa eta aldaketa
Suediera	
Rybing *et al.* (2010)	Esaldietan sintagmak ezabatu edo ordezkatu eta informazio sintaktikoa gehitu
Rennes eta Jönsson (2015)	Ordezkapena, ezabatzea, hurrenkera-aldaketa eta banaketa
Frantsesa	
Seretan (2012)	Ezabaketa, banaketa eta desenfatizazioa
Brouwers *et al.*; Brouwers *et al.* (2012; 2014)	Ezabaketa, modifikazioa eta banaketa
Bulgariera	
Lozanova *et al.* (2013)	Esaldiak banatu, esaldi konplexuak sinplifikatu, anafora ebatzi, subjektuak berreskuratu, perpausen hurrenkera zehaztu eta sintagma osagarriak txertatu
Koreera	
Chung *et al.* (2013)	Esaldiak banatu eta argumentuak tokiz aldatu

2.3 taula – Sinplifikazio sintaktikoa egiteko sinplifikazio-eragiketak

Sinplifikazio sintaktikoa egiten duten sistemen arkitekturak 2.4.2 atalean aurkeztuko ditugu eta 2.4.3 atalean sistemak teknikaren arabera sailkatuko ditugu.

Sinplifikazio lexikala

Azken urteotan sinplifikazio-mota honek hartu duen tokiaren adierazle dugu SemEval lehiaketan ataza bat antolatu izana (Specia *et al.*, 2012). Ataza horretan 5 sistemek hartu zuten parte eta teknika ezberdinak erabili zituzten (Amoia eta Romanelli, 2012; Jauhar eta Specia, 2012; Johannsen *et al.*, 2012; Ligozat *et al.*, 2012; Sinha, 2012). Sinplifikazio lexikalaren adibideak ingelesez

eta gaztelaniaz 2.4 taulan ikus daitezke.

Hizkuntza	**Jatorrizkoa**	**Sinplifikatutakoa**
Ingelesa Specia *et al.* (2012)	Hitler committed terrible *atrocities* during the second World War.	Hitler committed terrible *cruelties* during the second World War.
Gaztelania Bott *et al.* (2012a)	El visitante puede contemplar los óleos y esculturas que se exponen en la *pinacoteca*.	El visitante puede contemplar los óleos y esculturas que se exponen en el *museo*.

2.4 taula – Sinplifikazio lexikalen adibideak ingelesez eta gaztelaniaz

Bi metodo nagusi erabil daitezke lexikoa sinplifikatzeko: hiztegietan, datubase lexikaletan eta hitzen maiztasunetan oinarritzen dena ala estatistika erabiltzen duena. Lan gehienetan bien konbinazioak erabili dituzte. Gehien erabili diren baliabideak, berriz, *WordNet* (Fellbaum, 2010) datu-base lexikala eta *Wikipedia* izan dira. Bi horien erabileraren arabera sailkatuko ditugu aurkeztuko ditugun lanak.

WordNet erabiliz, sinplifikazio lexikala aztertzen duten hainbat lan aurki ditzakegu. De Belder *et al.*en (2010) metodoak bi motako hitz alternatiboen multzoak sortzea proposatzen du ordezkatu nahi den hitzari zuzenduak. Lehendabiziko hitz-multzoa sinonimoak dituen hiztegi batetik edo WordNetetik lortzen dute, eta bigarrena *Latent Words* hizkuntza-eredua erabilita. Amaierako hitzaren aukeraketa probabilitate bidez kalkulatzen dute hiru baliabide hauetan oinarrituta: psikolinguistikako neurriak dituen datubasea, testu errazen corpuseko unigramen probabilitatea eta silaba-kopurua. Biran *et al.*ek (2011) hitzen testuingurua kontuan hartzen duen bi mailako sistema proposatzen dute. Lehenengo mailak erregelak erauzten ditu eta bigarrenak sinplifikazioa egiten du. Erregelak erauztean, lehenik, sinplifikatzeko hautagaiak diren edukizko hitz guztientzat (*stop words*, zenbakiak eta puntuazio-markak baztertuz) bektore bana eraikitzen dute eta, ondoren, hitz bakoitza ordezkatuko duen hautagaiak lortzeko WordNet erabiltzen dute. Sinplifikatzean, jatorrizko esaldiaren testuinguruak bi modutan eragiten du: hitz-esaldien antzekotasuna eta testuinguruaren antzekotasuna. Egileen arabera sistema hori zazpi hitz baino gehiagoko esaldientzat da egokia. Thomas eta Anderson-ek (2012) sei algoritmo probatzen dituzte sinplifikazio lexikal egokiena lortzeko. Algoritmo horiek Personalized Page Rank eta informazio maximizazioaren printzipioak erabiltzen dituzte. Nunes *et al.*ek (2013) lau pausotan banatutako metodoa aurkezten dute: kategoriak etiketatzea,

sinonimoak identifikatzea, testuinguruaren maiztasunaren araberako ordezkapena eta esaldia zuzentzea. WordNetez gain, sinonimoen datu-base bat erabiltzen dute. Ordezkapenak egitean hitzen maiztasunak bilatzeko, haurren literaturako liburuekin osatutako hiztegi batean eta web bilatzaileetan oinarritzen dira.

Wikipedia erabiltzen duten lanen artean, berriz, beste hainbat daude. Ingelesez, *Simple Wikipedia*ko edizioen historiala erabiliz Yatskar *et al.*ek (2010) bi hurbilpen proposatzen dituzte: i) edizio-eragiketa guztiekin modelo probabilistiko bat sortzen dute eta ii) sinplifikazioak ez diren errebisioak iragazteko metadata erabiltzen dute. Ligozat *et al.*ek (2013) lexikoa sinplifikatzean kontuan hartzeko hiru irizpide aurkezten dituzte eta bakoitzari dagokion eredua aurkezten dute, *Simple Wikipedia*ko terminoen frekuentziak erabiliz, n-grametan oinarrituz eta kookurrentzien informazioa hartuz. Bayes teoreman oinarrituta, Shardlow-ek (2012) hitz bat testuinguru sinple batean agertzen den probabilitatea kalkulatzen du hitzen frekuentziaren gaineko kontaketak *Wikipedia* eta *Simple Wikipedia*tik hartuz. Hitzak ordezkatzeko erabiltzen duen baliabide lexikala WordNet da. Shardlow-ek (2013a) sinplifikatzeko hautagaiak diren hitz konplexuak identifikatzeko metodoak azaltzen ditu. Kauchak-ek (2013) hizkuntza-eredu bat egokitzean, sinplifikatu gabeko datuak (datu normalak) erabiltzearen eragina aztertzen du eta ondorioztatzen du datu normal gehigarriek ingeles errazaren hizkuntza-ereduen performantzia hobetzen dutela.

Suedieraz, sinonimoen ordezkapenaren bitartez egiten du Keskisärkkäk (2012). Ordezko sinonimoak aukeratzeko hiru estrategia erabiltzen ditu: hitzen maiztasuna, hitzen luzera eta sinonimia.

Gaztelaniaz, LExis sistema eraiki dute Bott *et al.*ek (2012a). Lan horretan *OpenThesaurus* baliabide lexikalean oinarritzen dira eta ordezkapenerako hautagai hoberena aukeratzen dute hitzei sinpletasun neurri bat eman ondoren eta teknika ezberdinak probatuz. WordNet eta *OpenThesaurus*en arteko konbinazioarekin ere esperimentuak egin dituzte (Saggion *et al.*, 2013). Hitzen testuingurua kontuan hartzen duen algoritmo baten bitartez emaitzak hobetzea lortzen dute Baeza-Yates *et al.*ek (2015) eta hitzen desanbiguazioa egiteko estrategiak ere erkatzen dituzte Saggion *et al.*ek (2016)).

Japonieraz, datu-multzo bat eta sistema bat aurkezten dute Kajiwara eta Yamamoto-k (2015). Sistema horrek hitz konplexuak identifikatzen ditu, ordezkapenak sortzen ditu, hitzen desanbiguazioa egiten du eta sinonimoen ranking batean hitz sinpleena aukeratzen du.

Azpiatal honekin bukatzeko, aipatu nahi dugu badirela ere kode irekia du-

ten eta sinplifikazio lexikala egiteko prest dauden inguruneak. LEXensteinek (Paetzold, 2015; Paetzold eta Specia, 2015) ordezkapenak sortzen ditu, ordezkapena aukeratzen du, ordezkapenak ranking batean jartzen ditu eta ezaugarriak estimatzen ditu. Ezaugarri horiek lexikalak, morfologikoak, kolokazionalak eta esanahiari orientatuak dira.

Bestelako sinplifikazioak eta domeinutara egokitutako sinplifikazioak

Sinplifikazio lexikala edo sintaktikoa ez diren lanak izango ditugu hemen aztergai. Adibidez, zenbakizko adierazpenen birformulazioa (Bautista *et al.*, 2013) edo sinplifikazio semantikoa (Kandula *et al.*, 2010).

Zenbakizko adierazpenen birformulazioen sinplifikazioa gauzatzeko, esaterako, gaztelaniazko *1,9 millones de hogares* katean zenbakizko adierazpenak eta lexikoa sinplifikatu ondoren, *2 millones de casas* lortzen da (Bautista *et al.*, 2012a; Bautista eta Saggion, 2014). Zenbakizko adierazpenak sinplifikatzeko bost eragiketa aurkezten dituzte: parentesien arteko zenbakiak ezabatzea, hizkiz dauden zenbakiak zifraz ematea, kantitate handiak hitzen bitartez adieraztea, biribiltzea eta hamarrekoak ezabatuz biribiltzea.

Sinplifikazio semantikoaren atazan, azalpenak gehitzen dira; adibidez, ingelesezko *Humerus* hitzari azalpena parentesi artean gehitzean, *Humerus (a part of arm)* lortzea da (Kandula *et al.*, 2010).

Diziplinei edo jakintza-alorrei dagokienez, TSAn bi domeinu izan dira bereziki landu direnak: biomedikuntza eta larrialdien kudeaketa. Biomedikuntza-domeinuan helburu-taldeak HPko aplikazioak eta gizakiak izan dira; larrialdien kudeaketan, berriz, gizakiak.

Horrela, bada, biomedikuntzako literatura sinplifikatzeko edo osasun-informazioa gizakiei ulerterrazagoa egiteko SIMTEXT sistemak (Damay *et al.*, 2006; Ong *et al.*, 2007) lexikoan sinonimoak ordezkatzen ditu eta sintaxian perpausak banatzeko erregelak eta transformazio-erregelak. Medikuntzako lexikoa sinplifikatzeko, Leroy *et al.*ek (2013) ordezkapen-hitzak proposatzen dituen algoritmo bat testu-editore batean integratzeko helburua dute, ondoren aditu batek balidazio- edo gainbegiratze-urratsean hitzik egokiena aukera dezan gramatikaltasuna bermatzearekin batera. Algoritmo horrek bi pausotan egiten du lan: lehenik, termino zailak identifikatzen ditu *Google Web Corpus*ean[11], n-grama-kontaketak eginez eta agerpen gutxi dituztenak hitz

[11] http://catalog.ldc.upenn.edu/LDC2009T25 (2013ko urrian atzituta)

zailagoak direla onartuz. Ondoren, termino horien hitz alternatibo errazak proposatzen ditu WordNeteko sinonimoak eta hiperonimoak erabiliz, definizioak eta mota semantikoak *Unified Medical Language System*etik[12] (UMLS), eta definizioak ingelesezko *Wiktionary*tik[13] eta *Simple Wiktionary*tik[14]; soilik kategoria gramatikal bera duten hitzak proposatzen ditu.

HPko aplikazio aurreratuak helburu dituzten lanekin jarraituz, biomedikuntzako artikuluen laburpenetako analizatzaile sintaktikoaren emaitzak hobetzeko eta testu horietako informazioa erauzteko algoritmoa garatu dute Jonnalagadda *et al.*ek (2009). Algoritmo horrek Link Grammar analizatzaile sintaktikoaren bitartez hitz-pareen arteko erlazio gramatikal jakinak eta puntuazio-markak erabiltzen ditu; horrela, esaldiak perpausetan banatzen dituzte. BioSimplify sisteman izen-sintagmen ordezkatzeak ere gehitzen dituzte Jonnalagadda eta Gonzalez-ek (2010a). Proteinak erauztea helburu duen esperimentuan (Jonnalagadda eta Gonzalez, 2010b), Siddharthan-en (2006) azaldutako erregelak aplikatu ondoren, sistemak lortzen diren esaldiak gramatikalak diren ala ez egiaztatzen du eta gramatikalak direnak aukeratzen ditu.

Koordinazioaren fenomenoa analizatuz, Evans-ek (2011), esaldi koordinatuen etiketatze sakonean oinarrituta, lau sailkatzaile probatzen ditu informazioa erauzteko helburuarekin. Esaldiak berridazteko, corpus-azterketan oinarritutako eskuzko erregelak erabiltzen ditu.

iSimp sistemak (Peng *et al.*, 2012) biomedikuntzako testu zientifikoen laburpenak sinplifikatzen ditu testu-meatzaritza egiteko helburuarekin, patroiak erabiliz. Formatuan aldaketak eginez, Peng *et al.*ek (2014) sisteman hobekuntzak lortzen dituzte.

Biomedikuntzako testuetan gertaeren erauzketa helburu izanik, Minard *et al.*ek (2012) ataza horretarako beharrezkoa den informazioarekin geratzeko sinplifikatzen dituzte esaldiak. Horretarako, corpus txiki baten etiketatzean oinarrituta, CRF (*Conditional Random Fields*) sailkatzailea erabiltzen dute esaldiak etiketatzeko. Esaldi horiek informazioa erauzteko euskarri bektoredun makinen sarrera izango dira. Esaldiak etiketatu ahala, corpusa handitzen doaz.

Itzulpen automatikorako testuinguru baten barnean, Medikuntza arloko testuak ingelesetik txinerara itzultzen dituen sisteman, Chen *et al.*ek (2012)

[12] http://www.nlm.nih.gov/research/umls/ (2013ko irailean atzituta)

[13] http://en.wiktionary.org/wiki/Wiktionary:Main_Page (2013ko irailean atzituta)

[14] http://simple.wiktionary.org/wiki/Main_Page (2013ko irailean atzituta)

eskuzko erregelen bitartez sinplifikatzen dute sintaxia eta lexikoa terminoak ordezkatuz.

Larrialdien kudeaketaren domeinuetako testuak ulerterrazagoak egiteko, Temnikova *et al.*ek (2012) hizkuntza kontrolatu bat proposatzen dute, eta instrukzioak dituzten testuak sinplifikatzeko, bost motatako erregelak ematen dituzte: orokorrak, formatuaren gainekoak, sintaktikoak, lexikalak eta puntuazio-marken gainekoak. Horretaz gain, sinplifikatutako testuek egitura jakin batzuk jarraitu behar dituztela zehazten dute larrialdi-kasuetan eraginkorrak izan daitezen: izenburua, azpitituluak, baldintzak, egin behar diren ekintzak (instrukzioak), oharrak (azalpenak) eta zerrendatzeak. Izenburuak eta instrukzioak ezinbestean azaldu behar badute ere, beste elementuak aukerakoak dira.

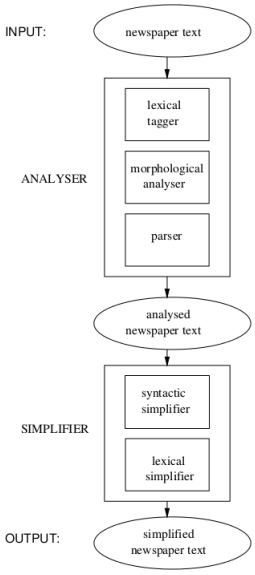
2.4.2 TSAko sistemen arkitekturak

TSAn aurkeztu zen lehendabiziko lanean (Chandrasekar *et al.*, 1996) sinplifikazio sintaktikoa erregela bidez lantzen zen, eta analisi linguistikoan oinarritzen zen (*chunk*etan eta dependentzietan). Hurrengo lanean, Chandrasekar eta Srinivas-ek (1997) une bakoitzean esaldi bana prozesatzen duen bi mailako arkitektura aurkezten dute: analisia eta transformazioak. Analisiak osagaien eta dependentzien informazioa erabiltzen du, eta transformazioak egiteko, erregelak automatikoki erauztea proposatzen dute domeinuetara errazago egokitzeko.

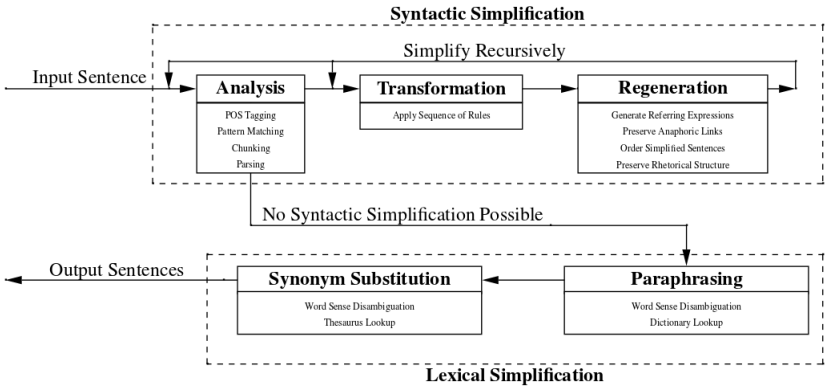
PSET proiektuan, Carroll *et al.*ek (1998) ere bi mailako arkitektura proposatzen dute: analizatzailea eta sinplifikatzailea. Analizatzaileak hiru modulu ditu: lexiko-etiketatzailea, analizatzaile morfologikoa eta analizatzaile sintaktikoa. Sinplifikatzaileak, berriz, bi modulu dauzka: sintaxi-sinplifikatzailea eta lexiko-sinplifikatzailea. Sistema horren arkitektura 2.2 irudian ikus daiteke.

TSAn eragin handia izan duen lana Siddharthanena da, sistemen oinarrizko arkitektura finkatu baitzuen (Siddharthan, 2002) (2.3 irudia). Sistema horrek hiru mailako exekuzio-hodia du sintaxia sinplifikatzeko, eta bi mailakoa lexikoa sinplifikatzeko. Sintaxiko hiru maila horiek analisia (*analysis*), transformazioa (*transformation*) eta birsorkuntza (*regeneration*) dira, eta lexikokoak parafrasiak (*paraphrasing*) eta sinonimoen ordezkapenak (*synonym substitution*).

Sinplifikazio sintaktikoan, Siddharthan-ek (2006), transformazio-moduluan erregelen hurrenkera gehitzen du eta birsortze-moduluan bi ataza na-



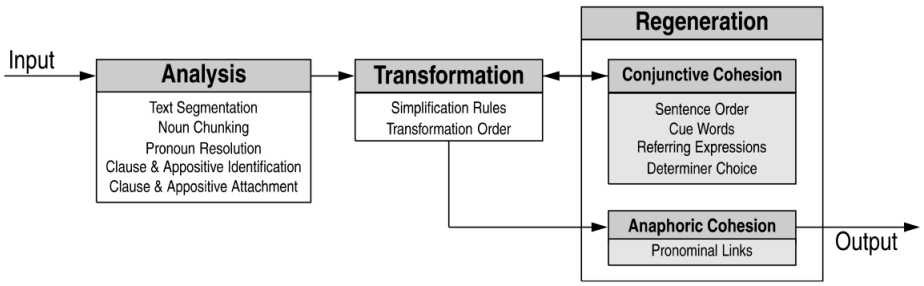
2.2 irudia – Sistemaren arkitektura (Carroll *et al.*, 1998)



2.3 irudia – Sistemaren arkitektura (Siddharthan, 2002)

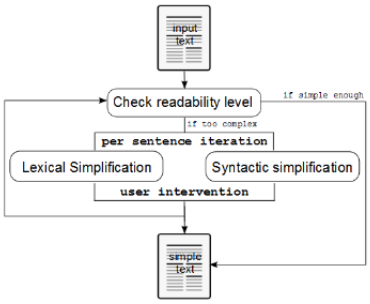
gusi azaltzen ditu: konjukzio bidezko kohesioa (esaldien arteko hurrenkera, txertatze-elementuaren aukeraketa, determinatzailearen aukeraketa) eta kohesio anaforikoa. Azkeneko modulu horren helburua testu berriaren kohesioa bermatzea da, eta horrekin, testu sinplifikatuen kohesioari garrantzia ematen zaio lehen aldiz. Sistemaren arkitektura 2.4 irudian ikus daiteke.

Arkitektura hori beste hizkuntza batzuetako lanen oinarria izan da.



2.4 irudia – Sistemaren arkitektura (Siddharthan, 2006)

Brasilgo Portugeseko SIMPLIFICA sistemak (Scarton *et al.*, 2010) hiru modulu nagusi ditu: konplexutasunaren ebaluazioa (Aluísio *et al.*, 2010), sinplifikazio sintaktikoa (Candido *et al.*, 2009) eta sinplifikazio lexikala. Sinplifikazio sintaktikoaren modulua Siddharthan-en (2006) lanean oinarritzen da. Sinplifikazioa gauzatu ondoren, erabiltzaileek sinplifikatutako esaldirik onena aukera dezakete. Sistema horren arkitektura 2.5 irudian ikus daiteke.



2.5 irudia – SIMPLIFICA sistemaren arkitektura (Scarton *et al.*, 2010)

Arabierarako proposatu den Al-Baseet sistemak (Al-Subaihin eta Al-Khalifa, 2011) lau moduluko arkitektura du: konplexutasunaren ebaluazioa, sinplifikazio lexikala, sinplifikazio sintaktikoa eta arabiar hizkuntzaren tipologia dela-eta diakritizazioa.

Lehen lan horiek erakusten duten bezala, sintaxia sinplifikatzeko hasierako arkitekturetan bi modulu (analizatzailea eta transformatzailea edo sinplifikatzailea) agertzen dira. Urteak aurrera joan ahala, modulu gehiago gehitu

izan zaizkie sinplifikazio-mota ezberdinak lantzeko, testuen konplexutasuna ebaluatzeko edo lantzen dituzten hizkuntzei lotutako behar bereziak tratatzeko.

2.4.3 TSAko sistemak teknikaren arabera

Atal honetan, testuen sinplifikazio automatikoa egiten duten sistemak sailkatu ditugu erabiltzen duten teknikaren arabera. Teknika horiek ezagutza linguistikoan oinarritutakoak, estatistikan oinarritutakoak, itzulpen automatiko-estatistikoan oinarritutakoak eta teknika hibridoetan oinarritutakoak dira. Atal honetan aurkeztuko ditugun sistemek sinplifikazio sintaktikoa edo bi sinplifikazioak (sintaktikoa eta lexikoa) egiten dituzte.

Ezagutza linguistikoan oinarritutako sistemak

Ezagutza linguistikoan oinarritzen diren sistemetan hizkuntzalariek edo adituek testuak sinplifikatzeko erregeletan kodetzen dute informazioa. Erregela horiek definitzeko, egile bakoitzak bere formalismoak erabili ditu eta, normalean, erregela horiek erabiltzen dituzten analizatzaileen irteeretara egokituta daude. Ondoren, motor batek erregela horiek interpretatzen ditu. Erregela horiek definitzeko informazioa corpusak aztertuz lortu dute, edo helburutalde jakin bat duten sistemetan, helburu-talde horren beharrei erantzuteko adituek definitu dituzte.

Atal honi, erdal hizkuntzetan garatu diren sistemez gain, gizakiak helburu dituzten lanekin emango diogu hasiera. Ondoren, HPko tresnak helburu dituzten sistemak aipatuko ditugu.

Chandrasekar *et al.*en (1996) lanean, *chunk*etan eta dependentzia-zuhaitzetan oinarritzen diren hurbilpenak erkatzen dira. Hurrengo lanean, erregelak ikasketa automatikoaren bitartez erauztea proposatzen dute Chandrasekar eta Srinivas-ek (1997).

PSET proiektuaren helburua afasikoei egunkariko testuak irakurterrazagoak egitea da (Carroll *et al.*, 1998). Sistema horretan, sintaxi-sinplifikatzaileak esaldi pasiboen aktiborako bihurketa, txertatutako perpausen erauzketa eta esaldien banaketa tratatzen ditu. Lexiko-sinplifikatzaileak WordNeteko sinonimoak hartu eta Oxfordeko Psikolinguistikako datu-basean galdetuz, sinonimo bakoitzaren maiztasunak lortzen ditu. Ondoren, sinplifikazio-mailaren arabera, sinonimo bat edo beste aukeratzen du. Sistema hori afasikoek dituzten beharrak kontuan hartuta eraiki dute. Hurrengo lanean,

Canning eta Tait-ek (1999) sintaxia sinplifikatzeko sistema azaltzen dute eta horrek hiru moduluetan honako atazak egiten ditu: i) anaforaren ebazpena, ii) sinplifikazioa eta iii) lehendabiziko moduluak detektatutako izenordainei dagozkien izen-sintagmen ordezkapena.

Siddharthan-en (2002) sistemak helburu-taldea irekia du eta bertan sinplifikazioa gauzatzeko arkitektura aurkezten du. Sintaxia sinplifikatzeko arkitektura Siddharthan-en (2006) zehazten du. REGENT sistema, berriz, (Siddharthan, 2011) dependentzia motatuetan oinarritzen da eta 63 erregelaz osatzen da. Esaldiak sortzeko bi aukera dauzka: i) *gen-light*: transformatutako dependentzia-grafoak hitzen hurrenkerarekin eta jatorrizko esaldiaren morfologiarekin lerratzea eta ii) *gen-heavy*: Stanfordeko dependentziak DSyntS errepresentazio bihurtu ondoren, esaldiak RealPro azalerako errealizatzailearekin sortzea. Ondorengo lan batean, sinplifikazio lexikala gehituz, Angrosh eta Siddharthan-ek (2014) erakusten dute automatikoki lortutako sinplifikazio lexikaleko erregelak nola orokor daitezkeen eta sistema hibrido batean sintaxia sinplifikatzeko automatikoki eta eskuz idatzitako erregelak nola konbina daitezkeen. Dependentzia-gramatika sinkronikoak erabiltzen dituzte oinarri gisa.

Max-ek (2006), gaixotasun kognitiboak dituzten pertsonei (batez ere, afasikoei) testuak egokitzen dizkieten egileei laguntzeko, sinplifikazio-erregelak testu-editore batean integratzen ditu. Erregela horiek dependentzia-zuhaitzetan oinarritzen dira eta hizkuntzalariek idatzi dituzte.

Ingelesa aztergai ez duten beste hizkuntzetako sistemak aipatuko ditugu orain. Japonierarako eta jaiotzetiko entzumen arazoak dituzten pertsonei zuzenduta, 28.000 erregela baino gehiago inplementatu dituzte Inui *et al.*ek (2003). Erregela horietan parafrasi lexikalak (sinonimoen ordezkapena) eta parafrasi sintaktikoak egiten dituzte. Ondoren, horien postedizioa egitea proposatzen dute.

Brasilen alfabetatze-arazo larriak dituztenez, testuen ulermena areagotzeko, PorSimples proiektua sortu zen. Proiektu horretan Brasilgo portugesean linguistikoki zailak diren fenomenoak aztertu dituzte eta sinplifikazio-erregelak definitu dituzte Aluísio *et al.*ek (2008b). Sistema horrek sinplifikatu aurretik testuak konplexuak diren ala ez ebaluatzen du (Gasperin *et al.*, 2009a). Sistema horri sinplifikazio lexikala gehituz, SIMPLIFICA sistema (Scarton *et al.*, 2010) sortu dute, testu sinplifikatzailea integratuta duen testu-editorea. Lexikoa sinplifikatzeko, hitz konplexuak eta sinpleak dituzten hiztegietan oinarritzen dira.

Gaztelaniako testuak arazoak dituzten gizakiei irisgarriagoak egiteko, Sim-

plext proiektua[15] (Saggion *et al.*, 2011) abiatu zen. Proiektu horren barnean sinplifikazio sintaktikoa eta lexikala landu dira. Baliabideen arteko konparazioak ere egin dira. Sintaxia sinplifikatzeko sistemak hiru pauso ditu: i) gramatika batek sinplifikatu behar diren egiturak bilatu eta etiketatzen ditu; ii) iragazki estatistiko batek berresten du ea benetan etiketatutako esaldiak sinplifikatu behar diren ala ez eta iii) manipulazio sintaktikoak egiten ditu. Sistema horretan sinplifikazio lexikala eta materia-sinplifikatzailea integratzeko asmoa dutela adierazten dute. Sistema hori corpus-azterketa batean oinarritzen da. Drndarević *et al.*en (2013) lanean sintaxia eta lexikoa sinplifikatzen dituzten moduluak ebaluatzen dituzte.

Azken lanetan, aurreko lanak osatuz, sistema modularra aurkeztu dute Saggion *et al.*ek (2015b). Sistema horrek hiru modulu ditu: i) sinplifikazio sintaktikoko osagaia, ii) sinonimoetan oinarritutako sinplifikazio osagaia (semantika distribuzionalean oinarrituz) eta iii) erregeletan oinarritutako sinplifikazio lexikaleko osagaia. Analisirako eta erregelak aplikatzeko, dependentzia-zuhaitzetan oinarritzen dira.

Suedieraz sinplifikazio sintaktikoa egiteko, berriz, CogFLUX sistemak (Rybing *et al.*, 2010) irakurketa errazeko eta testu normaletako corpus-azterketan oinarritutako 25 transformazio-erregela erabiltzen ditu. Sistema horri sinplifikazio-eragiketa berriak gehitu dizkiete Rennes eta Jönsson-ek (2015).

Frantseseko egitura sintaktikoak sinplifikatzeko erregelak erauztean erabiltzen den eskuzko metodoa osatzeko, Seretan-ek (2012) estatistikoki nabarmentzen diren egitura linguistikoak proposatzen ditu etiketatzaileei laguntzeko. Brouwers *et al.*ek (2012) 19 erregeletan oinarritutako sistema aurkeztu dute. Corpus-azterketetan oinarritzen dira erregela horiek definitzeko, eta sinplifikazioen tipologia proposatzen dute. Sinplifikazio onena zein den erabakitzeko, programazio lineal osoa erabiltzen dute (Brouwers *et al.*, 2014).

Italierarako, haurrek ipuinetako gertaerak hobeto uler ditzaten, ERNESTA *(Enhanced Readability through a Novel Event-based Simplification Tool)* izeneko sistema garatu dute Barlacchi eta Tonelli-k (2013)). Anafora ebatzi ondoren, eta gertaerak kontuan izanda, esaldiak sinplifikatzen dituzte informazio psikolinguistikoan oinarrituz.

Bulgarierarako, eta entzumen-arazoak dituztenak helburu izanik, Lozanova *et al.*ek (2013) multzo ezberdinetan (anaforaren ebazpena, perpausmugak, subjektuen berreskurapena, etab.) banatzen diren 23 erregeletako sistema aurkezten dute.

[15] `http://www.simplext.es/` (2015eko azaroan atzituta)

Gorrei koreerazko albisteen ulermena errazteko, Chung *et al.*ek (2013) esaldi laburragoak eta sinpleagoak sortzeaz gain, errepresentazio grafikoak erabiltzen dituzte.

Alemanerako ere sinplifikazio sintaktikoa egiten duen sistema aurkeztu dute (Suter, 2015).

HPko tresnak helburu dituzten sistemei dagokienez, tresnek informazioa eraginkortasun handiz prozesa dezaten, Beigman Klebanov *et al.*ek (2004) *Easy Access Sentences* kontzeptua proposatu dute. Sortzen diren esaldi horiek jatorrizko testuaren informazioa mantendu behar dute, eta aditz jokatu bakarraz eta entitate batez osatuak izan behar dira. Esaldi horiek, halaber, gramatikalak izan behar dira. Horrelako esaldiak erabilita tresnek errazago aurkitzen eta prozesatzen dute informazioa.

Azpitituluak egitean esaldiak sinplifikatzeko Daelemans *et al.*ek (2004) bi hurbilpen aurkezten dituzte, bata aurrerago azalduko dugun ikasketa automatikoan oinarritutakoa eta bestea erregeletan oinarritutakoa. Azken honetan nederlandera eta ingelesa sinplifikatzeko erregelak konpilatzen dituzte. Bi faseetan egiten dute sintagmen ezabatzea: i) erredundanteak diren esaldiak aukeratzen dituzte eta ii) trinkotasun-maila bateko esaldiak ezabatzen dituzte. Ezabatzeko hautagaiak adberbioak, adjektiboak, izen propioak, preposizio-sintagmak, egitura parentetikoak, erlatibozko perpausak, zenbakiak eta denbora-adierazpenak dira.

Ahozko hizkuntza ulertzen duten sistemen errendimendua hobetzeko, Tur *et al.*ek (2011) esaldiak sinplifikatzen dituzte. Beraien helburua sailkatzaile bat denez, eta ez gizakiak, sortzen dituzten esaldiek ez dute zertan gramatikalak izan.

Laburpen automatikoak egiteko, Bawakid eta Oussalah-ek (2011) Tregex patroiak erabiliz esaldiak banatzeko erregelak aplikatzen dituzte. Algoritmo baten bitartez esaldiak trinkotzen dituzte (*sentence compression*) laburpena egin aurretik.

Corpus paraleloen hitzak lerratzeko, gako-hitzetan oinarritutako zerrendak erabiliz banatzen dituzte esaldiak Srivastava eta Sanyal-ek (2012). Zerrenda horiek ingeleserako eta hindirako osatu dituzte.

Estatistikako tekniketan oinarritutako sistemak

Estatistikako teknikak erabiltzen dituzten sistemek helburu-talde irekia dute edo HPko aplikazio aurreratuei zuzenduta daude. Teknika estatistikoak askotarikoak dira eta ondoren aipatuko ditugu TSAn erabili direnak. Oro

har, teknika horietan corpusetatik informazioa induzitzen da eta bertatik ikasitakoa testuak sinplifikatzeko aplikatzen da.

Azpitituluak egiteko Daelemans *et al.*en (2004) bigarren hurbilpena ikasketa automatikoan oinarritzen da. Esaldiak sinplifikatzeko prozesua hitzen transformazio ataza bezala ulertzen dute; prozesu horretan, testuko hitzak kopiatu, ezabatu edo ordezkatu egiten dira.

Medero eta Ostendorf-ek (2011) testuen sinplifikazioan egindako aldaketa sintaktikoak identifikatu eta deskribatzen dituen sistema aurkezten dute ingelesa irakurterrazagoa izateko.

Bach *et al.*ek (2011) *log-linear* ereduan oinarritutako sistema eraiki dute. Metodo horrek ezaugarri sorta baten gainean lan egiten duen *margin-based discriminative learning* algoritmoa erabiltzen du. *Stack decoding* algoritmo batekin sinplifikazio-hipotesiak sortzen dituzte.

Itzulpenaren kalitatea hobetzeko, Doi eta Sumita-k (2004) esaldiak banatzen dituzte. Bi pausotan egiten dute: i) hautagaiak lortzeko n-grametan oinarritutako hizkuntza-ereduak (NLM, *N-gram Language Model*) erabiltzen dituzte, eta ii) hautagaien artean aukeratzeko, NLMa eta esaldien antzekotasuna erabiltzen dituzte.

Woodsend eta Lapata-k (2011a) jatorrizko eta helburu-testuak kontuan izanik, berridazketa konplexuak egiten dituzten erregelak ikasten dituzte. Hurbilpen hori *quasi-synchronous grammar*ean (QG) oinarrituta dago. Programa lineal oso bezala formulatuta dago eta QGa erabiltzen du berridazketa posible guztien espazioa estaltzeko. Egileen arabera, eredu hori kontzeptualki sinplea eta konputazionalki efizientea da.

Feblowitz eta Kauchak-ek (2013) zuhaitz sintaktikoen transformazioa egiten dute lerratutako corpus baten analisian probabilistikoki *synchronous tree substitution grammar*ekin (STSG) ordezkapenak eginez. Hauek dira jarraitzen dituzten pausoak: i) gramatika ikasi zuhaitzen osagaiak lerratuz, ii) gramatika osatu informazio lexikala gehituz, iii) egoera finituko transduktore bat aplikatu, entrenatutako *log-linear* eredu batekin *n-best* sinplifikazio hoberenen zerrenda osatzeko eta iv) puntuazio altuena duena aukeratu.

Paetzold eta Specia-k (2013) *Wikipedia*rekin ere zuhaitzen transformazioa egiten duten erregelak ikasten dituzte, bai sintaxia, bai lexikoa sinplifikatuko dituzten erregela bolumen handiak lortzeko. Metodo horrek hiru osagai nagusi ditu: entrenatzeko modulua, sinplifikazio modulua eta ranking modulua. Erregelak ikasteko Tree Transducer Toolkit (T3) erabiltzen dute.

Datu-kopuruak murritzak direnean, Vu *et al.*ek (2014) erregelen eskuzko berrikuspena proposatzen dute. Zein erregela aplikatu behar den jakiteko,

metodo bayesiarrak erabiltzen dituzte.

Hurbilpen ez-gainbegiratuak eta semantika erabiltzen dituzte Narayan eta Gardent-ek (2015). Ikasteko, *Wikipedia* eta *Simple Wikipedia* erabiltzen dute, baina lerrokatu gabe; esaldiak non banatu behar diren jakiteko, berriz, rol tematikoen antzekotasun maximoaren sekuentzian oinarritzen dira.

Danierarako, corpus handiak erabili gabe, Klerke eta Søgaard-ek (2013) azpiesaldien aukeraketa eginez esaldi sinpleak lortzen dituzte. Azpiesaldi hautagaiak lortzeko, dependentzia-gramatika baliatzen duen analizatzaile sintaktikoan oinarritzen diren heuristikoak ezabatzeko ausazko prozedura batekin konbinatzen dituzte. Heuristiko horiek esaldiaren osagaiak mantentzeko erabiltzen dituzte, eta hautagaien artean aukeratzeko funtzio-galera bat erabiltzen dute.

Itzulpen automatiko-estatistikoko tekniketan oinarritutako sistemak

Itzulpen automatiko-estatistikoko teknikak teknika estatitiskoen atalean sailkatu beharko liratekeen arren, azken urteetan teknika horrek hartu duen garrantzia dela-eta azpiatal propioa eskaintzea erabaki dugu. Ondoren aipatuko ditugun sistemek helburu-talde irekia dute.

Estatistikan oinarritutako itzulpen automatikoko (SMT, *statistical machine translation*) teknikak oinarri hartuta, zuhaitz sintaktikoen transformazioak egiten dituen eredua aurkezten dute Zhu *et al.*ek (2010). Eredua iteratiboki entrenatzeko, *expectation maximization* (EM) algoritmoa erabiltzen dute, eta entrenatze-prozesu hori bizkortzeko hizkuntza bakarreko hitzen mapaketan oinarritutako metodoa aplikatzen dute. Azkenik, dekodifikatzaile bat erabiltzen dute esaldi sinplifikatuak sortzeko estrategia irenskorrak (*greedy*) erabiliz eta hizkuntza-ereduak integratuz.

Coster eta Kauchak-ek (2011a) itzulpen automatikoko hurbilpenak erabiltzen dituzte, baina sintagmetan oinarritutako aldaera gehitzen dute. Itzulpen automatikoan aldaera horrek hobekuntzak lortu ez dituen arren, sintaxian oinarritutakoak baino emaitza hobeak lortu ditu. Wubben *et al.*ek (2012) ere esaldiak sinplifikatzeko sintagmetan oinarritutako itzulpen automatikoa erabiltzen dute, antzekotasun-ezan (*dissimilarity*) oinarritutako *re-ranking* heuristiko batekin areagotuz eta hizkuntza bakarreko corpus paralelo batekin entrenatuz.

Narayan eta Gardent-ek (2014), berriz, semantikan eta itzulpen automatikoan oinarritutako sistema aurkezten dute. Sarrera gisa, semantika sako-

naren errepresentazioa hartzen dute eta sintagma-mailako itzulpen automatikoko eredu batean oinarritzen dira banaketak eta ezabaketak egiteko.

Teknika horiek gaztelaniaz eta portugesez ere erabili izan dira. Speciak (2010) itzulpen automatiko-estatistikoko teknikak erabiltzen ditu Brasilgo portugeserako. Metodo horrekin emaitza onak lortzen ditu batez ere sinplifikazio lexikalean eta berridazketa sinpleetan. Gaztelaniarako ere, itzulpen automatikoko metodoak proposatu dituzte, eta sinplifikazio leuntzat egokia dela ondorioztatzen dute Štajner-ek (2014) modelo ezberdinak probatu ondoren (Štajner *et al.*, 2015b). Sintagmetan oinarritutako itzulpen automatikoko hiru hurbilpen hiru hizkuntzatan (gaztelania, Brasilgo portugesa eta ingelesa) probatu dituzte Štajner eta Saggion-ek (2015). Esaldiak oso antzekoak direnean funtzionatzen duela aipatzen dute.

Teknika hibridoetan oinarritutako sistemak

Teknika hibridoetan oinarritzen diren sistemak, oro har, erregeletan oinarritzen dira, baina ondoren estatistikako teknikak erabiltzen dituzte erregelen aplikazio edo sinplifikazio onena erabakitzeko. Sistema horiek HPko tresnak eta gizakiak dituzte helburu.

Rol semantikoak etiketatzeko, esaldia banatzen dute Vickrey eta Kollerek (2008). Horretarako, eskuzko erregela sintaktikoak aplikatzen dituzte eta ondoren ikasketa automatikoaren bidez zein erregela aplikatu erabakitzen dute.

Haurrei zuzendutako sistema garatzeko, lexikoa sinplifikatzeko WordNetetik lortutako hitz alternatiboak hizkuntza-eredu batekin konbinatzen dituzte De Belder eta Moens-ek (2010). Sintaxian egiturak erregelen bidez sinplifikatzen dituzte eta ondoren programazio lineal osoaren bitartez testuan, oro har, erregelek izan dezaketen eragina kalkulatzen dute sinplifikazio hoberena aukeratzeko. Testuak sinplifikatu aurretik, testuen konplexutasuna ere aztertzen dute.

Atal honetan TSAn egin diren sinplifikazio-motak aurkeztu ditugu eta batez ere sinplifikazio sintaktikoan lantzen diren fenomenoak eta eragiketak aipatu ditugu. Sinplifikazio-mota horiek egiten dituzten sistemen arkitekturak aurkeztu ditugu eta sistemek erabiltzen dituzten teknikaren arabera (ezagutza linguistikoan oinarritutakoak [Erreg.], estatistikako tekniketan oinarritutakoak [Estat.], itzulpen automatiko-estatistikoko tekniketan oinarritutakoak

[IA] eta teknika hibridoetan oinarritutakoak [Hibr.]) sailkatu ditugu. Laburbiltzeko asmoz, aurkeztu ditugun teknikaren eta sinplifikazio-motaren araberako ingelesezko sistemak 2.5 taulan jaso ditugu eta 2.6 taulan gainontzeko hizkuntzetako sistemak.

Ingeleseko sistemak	**Sinpl. sintaktikoa**				**Sinpl. lexikala**	**Best.**
	Erreg.	Estat.	IA	Hibr.		
Ingelesa Chandrasekar *et al.* (1996), Chandrasekar eta Srinivas (1997)	✔	-	-	-	-	-
PSET proiektua, Systar Carroll *et al.* (1998), Canning eta Tait (1999)	✔	-	-	-	✔	-
Siddharthan (2002), Siddharthan (2006), Siddharthan (2010), Siddharthan (2011), Angrosh eta Siddharthan (2014)	✔	-	-	-	✔	✔
Beigman Klebanov *et al.* (2004)	✔	-	-	-	-	-
Daelemans *et al.* (2004)	✔	✔	-	-	-	-
Doi eta Sumita (2004)	-	✔	-	-	-	-
Max (2006)	✔	-	-	-	-	
SIMTEXT Damay *et al.* (2006), Ong *et al.* (2007)	✔	-	-	-	✔	-
Vickrey eta Koller (2008)	-	-	-	✔	-	-
BioSimplify Jonnalagadda *et al.* (2009), Jonnalagadda eta Gonzalez (2010a), Jonnalagadda eta Gonzalez (2010b)	✔	✔	-	-	-	-
De Belder eta Moens (2010), De Belder *et al.* (2010)	-	-	-	✔	✔	-
Kandula *et al.* (2010)	-	-	-	-	-	✔

(Jarraipena hurrengo orrialdean)

Ingeleseko sistemak	**Sinpl. sintaktikoa**				**Sinpl. lexika-la**	**Best.**
	Erreg.	Estat.	IA	Hibr.		
Yatskar *et al.* (2010)	-	-	-	-	✔	-
Zhu *et al.* (2010)	-	-	✔	-	✔	-
Bach *et al.* (2011)	-	✔	-	-	-	-
Bawakid eta Oussalah (2011)	✔	-	-	-	-	-
Biran *et al.* (2011)	-	-	-	-	✔	-
Coster eta Kauchak (2011a)	-	-	✔	-	✔	-
Evans (2011)	-	-	-	✔	-	-
Medero eta Ostendorf (2011)	-	✔	-	-	-	-
Poornima *et al.* (2011)	✔	-	-	-	-	-
Tur *et al.* (2011)	✔	-	-	-	-	-
Woodsend eta Lapata (2011a)	-	✔	-	-	✔	-
Amoia eta Romanelli (2012)	-	-	-	-	✔	-
Chen *et al.* (2012)	✔	-	-	-	✔	-
Jauhar eta Specia (2012)	-	-	-	-	✔	-
Johannsen *et al.* (2012)	-	-	-	-	✔	-
Ligozat *et al.* (2012), Ligozat *et al.* (2013)	-	-	-	-	✔	-
Minard *et al.* (2012)	-	-	-	-	✔	-
iSimp Peng *et al.* (2012), Peng *et al.* (2014)	✔	-	-	-	-	-
Shardlow (2012), Shardlow (2013a)	-	-	-	-	✔	-
Silveira Botelho eta Branco (2012)	✔	-	-	-	-	-
Sinha (2012)	-	-	-	-	✔	-
Specia *et al.* (2012)	-	-	-	-	✔	-
Srivastava eta Sanyal (2012)	✔	-	-	-	-	-
Temnikova *et al.* (2012)	✔	-	-	-	✔	-

(Jarraipena hurrengo orrialdean)

Ingeleseko sistemak	**Sinpl. sintaktikoa**				**Sinpl. lexikala**	**Best.**
	Erreg.	Estat.	IA	Hibr.		
Thomas eta Anderson (2012)	-	-	-	-	✔	-
Wubben *et al.* (2012)	-	-	✔	-	✔	-
Bautista *et al.* (2013)	-	-	-	-	-	✔
Feblowitz eta Kauchak (2013), Kauchak (2013)	-	✔	-	-	✔	-
Leroy *et al.* (2013)	-	-	-	-	✔	✔
Nunes *et al.* (2013)	-	-	-	-	✔	-
Paetzold eta Specia (2013), Paetzold (2015), Paetzold eta Specia (2015)	✔	-	-	-	✔	-
Vu *et al.* (2014)	-	✔	-	-	-	-
Narayan eta Gardent (2014), Narayan eta Gardent (2015)	-	✔	✔	-	✔	-
Štajner eta Saggion (2015)	-	-	✔	-	✔	-

2.5 taula – TSAko ingeleseko sistemak sinplifikazio-motaren eta teknikaren arabera

Hizkuntzak eta sistemak	**Sinpl. sintaktikoa**				**Sinpl. lexikala**	**Best.**
	Erreg.	Estat.	IA	Hibr.		
Japoniera						
Inui *et al.* (2003)	✔	-	-	-	✔	-
Kajiwara eta Yamamoto (2015)	-	-	-	-	✔	-
Portugesa						
PorSimples proiektua Aluísio *et al.* (2008b), Candido *et al.* (2009), Scarton *et al.* (2010)	✔	-	-	-	✔	-
Specia (2010)	-	-	✔	-	✔	-
Silveira Botelho eta Branco (2012)	✔	-	-	-	-	-

(Jarraipena hurrengo orrialdean)

Hizkuntzak eta sistemak	**Sinpl. sintaktikoa**				**Sinpl. lexikala**	**Best.**
	Erreg.	Estat.	IA	Hibr.		
Štajner eta Saggion (2015)	-	-	✔	-	✔	-
Suediera CogFLUX Rybing *et al.* (2010), Rennes eta Jönsson (2015)	✔	-	-	-	-	-
Keskisärkkä (2012)	-	-	-	-	✔	-
Arabiera Al-Subaihin eta Al-Khalifa (2011)	✔	-	-	-	✔	-
Gaztelania Simplex proiektua Saggion *et al.* (2011) Bott *et al.* (2012a), Bott *et al.* (2012b), Saggion *et al.* (2013), Drndarević *et al.* (2013), Štajner (2014), Štajner *et al.*, 2015b, Štajner eta Saggion (2015), Saggion *et al.* (2015b), Baeza-Yates *et al.* (2015), Saggion *et al.* (2016)	✔	-	✔	-	✔	-
Bautista *et al.* (2012a), Bautista eta Saggion (2014)	-	-	-	-	-	✔
Fajardo *et al.* (2013)	-	-	-	-	-	✔
Frantsesa Brouwers *et al.* (2012) Brouwers *et al.* (2014)	-	-	-	✔	-	-
hline Seretan (2012)	-	-	-	✔	-	-
Daniera Klerke eta Søgaard (2013)	-	✔	-	-	-	-
Italiera ERNESTA Barlacchi eta Tonelli (2013)	✔	-	-	-	-	-

(Jarraipena hurrengo orrialdean)

Hizkuntzak eta sistemak	Sinpl. sintaktikoa				Sinpl. lexikala	Best.
	Erreg.	Estat.	IA	Hibr.		
Bulgariera Lozanova *et al.* (2013)	✔	-	-	-	-	-
Koreera Chung *et al.* (2013)	✔	-	-	-	-	-

2.6 taula – TSAko beste hizkuntzetako sistemak sinplifikazio-motaren eta teknikaren arabera

2.5 TSA sistemak ebaluatzeko metodoak

Testuen sinplifikazioa egiten duten sistemen ebaluazioa nola egin TSAko komunitatean irekita dagoen galdera da. Ebaluazio hori egiteko, zehazki metrika edo metodoren bat proposatzen ez den bitartean, egile bakoitzak bere metodoa erabiltzen du. Atal honetan aurkeztuko ditugu orain arte erabili diren ebaluazio-metodoak. Metodorik erabilienak itzulpen automatikoan erabiltzen diren neurriak, konplexutasun-neurriak eta erabiltzaileei edo etiketatzaileei galdeketak egitea dira. Normalean egileek metodo bat baino gehiago erabiltzen dituzte sistemak ebaluatzeko.

Sistemen moduluz moduluko ebaluazioa egin izan da hasierako lanetan, analisia egiten zuen moduluari garrantzia emanez. Chandrasekar *et al.*ek (1996) egin duten ebaluazioan beren analisirako bi hurbilpenak (*chunk*etan eta dependentzietan oinarritutakoak) alderatu dituzte. Siddharthan-ek (2002) moduluz moduluko ebaluazioa egiten du, baina etorkizuneko lanetan bi metodo proposatzen ditu sistemaren errendimendua oro har ebaluatzeko. Bi metodo horiek dira intrinsekoa (*intrinsically*), erabiltzaileen ebaluazioa erabiltzen duena, eta estrintsekoa (*extrinsically*), bere errendimenduak beste sistema batean (analizatzaile sintaktikoa, itzultzaile automatikoa) duen eragina neurtzen duena.

2.5.1 Erabiltzaileen bidezko ebaluazioa

Erabiltzaileekin egindako ebaluazioekin hasiz, Carroll *et al.*ek (1998) beren sistema ebaluatzeko, irakurketa-esperimentuak egin dituzte ikusmen-arazoak

ez dituzten afasikoekin. Horrez gain, esaldien ulergarritasuna eta sistemaren baliagarritasuna aztertzeko, subjektuak elkarrizketatu dituzte.

Begi-mugimenduaren neurtzailea edo *eye-tracker*a erabilita, gaztelanian sinplifikazio lexikalak izan dezakeen eragina aztertu da. Neurtzaile horren bitartez begiaren eta buruaren arteko mugimendua eta begirada non kokatuta dagoen neurtu da. Alde batetik, Rello *et al.*ek (2013a) lexikoa sinplifikatzeko bi estrategia probatu dituzte: hitzak sinonimo errazagoekin ordezkatzea eta sinonimo errazagoak eskaintzea hitz konplexuarekin batera. Beste aldetik, Rello *et al.*ek (2013b) sinplifikazio lexikalaren eragina neurtu dute gaztelaniako aditzen parafrasien bitartez (*confiar* eta *tener confianza* bezalako aditz eta kolokazio pareak kontrastatuz). Horretaz gain, Rello *et al.*ek (2013c) zenbakizko adierazpenak letraz edo hitzen bitartez ematea aztertu dute. Hiru esperimentu horiek dislexia diagnostikatuta duten pertsonekin egin dituzte, baina datuak kontrastatzeko kontrol-talde bat ere osatu dute.

Ingeleseko jatorrizko hiztunak eta jatorrizko hiztunak ez direnen etiketatzeak erabili dituzte Yatskar *et al.*ek (2010) lexikoa sinplifikatzen duten sistema ezberdinak ebaluatzeko, sistemak helburu-talde jakinik ez duen kasuetan. Hala, sistema horietako bakoitzak sortzen dituen ehun hitz-pare eta ausazko beste ehun hitz-pare hartu dituzte eta etiketatzaileei eskatu zaie bikote horietako bakoitzean adierazteko zein den sinpleagoa, zein konplexuagoa, antzekoak diren, erlaziorik gabekoak diren, zalantza sorrarazten dien edo erabakitzeko zaila gertatu zaien.

Crowdsourcing bidezko ebaluazioek azken urteetan arrakasta izan dute. De Belder eta Moens-ek (2010) ebaluazioa *Wikipedia* (jatorrizkoa eta sinplea) entziklopediako eta *Literacyworks*[16] web-orriko testuekin, eta *Amazon's Mechanical Turk crowdsourcing*erako[17] web-zerbitzua erabiliz egin dute. Sinplifikazio lexikoa ebaluatzeko, ordezkapena zuzena den ala ez galdetu dute. eta sinplifikazio sintaxikoa ebaluatzeko, esaldiak zuzenak diren ala ez.

Mechanical Turk web-zerbitzuaren bitartez ere Leroy *et al.*ek (2013) Medikuntza arloko testuen sinplifikazio lexikala ebaluatzeko bi parametro neurtu dituzte: nabaritutako zailtasuna eta benetako zailtasuna. Lehenengoa neurtzeko, 1-5 bitarteko Likert eskala erabili dute, 1 oso erraza izanik eta 5 oso zaila. Bigarrena neurtzeko, hiru neurri erabili dituzte: ulermena neurtzeko testuarekin batera agertzen diren bost aukera anitzeko galdera, ikasketarako testurik gabeko beste zazpi aukera anitzeko galdera eta informazioaren

[16] http://literacynet.org/cnnsf/ (2013ko urrian atzituta)
[17] https://www.mturk.com/mturk/welcome (2013ko irailean atzituta)

oroimena neurtzeko, bi estaldura-galdera libre.

Sinpletasun-maila erabiltzaileen bidez ebaluatzea eraginkorra dela aipatzen dute Lasecki *et al.*ek (2015). Izan ere, 2.500 etiketatzetan ikusi dute erabiltzaileak zuzen asmatzeko gai direla.

Oro har, hiru dira kontuan hartzen diren parametroak: gramatikaltasuna, esanahiari eustea eta sinpletasuna. Parametro horiek 1-5eko eskalan neurtu izan ohi dira (Coster eta Kauchak, 2011a; Wubben *et al.*, 2012; Drndarević *et al.*, 2013; Saggion *et al.*, 2015b; Štajner *et al.*, 2015a).

2.5.2 Konplexutasun-neurrien bidezko ebaluazioa

Aurreprozesu bezala erabiltzen diren TKA sistemak ere erabili izan dira sinplifikatutako testu horien konplexutasun-maila baxuagoa den aztertzeko. Adibidez, Siddharthan-ek (2006) Flesch konplexutasun-neurria erabili du egunkarietako testuen sinplifikazioak ebaluatzeko.

Lehen aipatu dugun galdeketa-metodoarekin batera, konplexutasuna aztertzeko formulak ere erabili dituzte Drndarević *et al.*ek (2013). Formula horiek ausaz aukeratutako 100 testuri aplikatu dizkiete; testu horiek hiru mailatan sinplifikatu dira: lexikala, sintaktikoa eta biak batera. Galdeketetan, berriz, hogeita bost etiketatzaileri hiru galdera erantzuteko eskatu diete. Galdera horiek jatorrizko esaldien gramatikaltasunari, sinplifikatutako esaldien gramatikaltasunari eta jatorrizko esaldiaren eta sinplifikatutako esaldiaren esanahien arteko ezberdintasunei buruzkoak dira. Multzo bakoitzarentzat erdiko joera zein den jakiteko batezbestekoa eta mediana kalkulatu dituzte, eta aldakortasunaren indikatzaile bezala, maiztasunen banaketa. Štajner *et al.*ek (2015c) ere konplexututasun-neurriak aplikatu dituzte.

Temnikova *et al.*ek (2012) bi motatako ebaluazioa egin dute: intrintsekoa eta estrintsekoa. Intrintsekoa, konplexutasuna neurtzen duten neurrien bitartez egiten dute. Estrintsekoa, berriz, hiru modutan egiten dute: irakurmenaren ulermenean duen eragina, eskuzko itzulpenean eta itzulpen automatikoan duen eragina eta amaierako erabiltzaileen onargarritasuna aztertu dituzte. Hirurak erabiltzaileekin egin dituzte; lehenengoa eta hirugarrena galdeketa bidez, eta bigarrena postedizio-esfortzua automatikoki neurtuz (denbora, ikuspuntu teknikoa eta ikuspuntu kognitiboa).

2.5.3 Itzulpen automatikoko neurrien bidezko ebaluazioa

Testuen sinplifikazio automatikoa itzulpen prozesu bat bezala uler daitekeenez, itzulpen automatikoko sistemak bezala ere ebaluatu izan da, itzulpen automatikoan (IA) erabiltzen diren metrikak aplikatuz.

Daelemans *et al.*ek (2004) ebaluazioak duen zailtasunari erreparatuta eta jakinda oso garestia dela eskuz ebaluatzea, BLEU metrika proposatu dute. Ildo horretatik jarraituz, Zhu *et al.*ek (2010) ere IAko BLEU and NIST neurriak erabili dituzte beren sistemaz gain sortu dituzten beste lau oin-lerro sistema ebaluatzeko.

Specia-k (2010) ere neurri horiek erabiltzeaz gain, kate-parekatzea eta eskuzko ebaluazioa egin ditu. Horrela, segmentuak egokiak eta naturalak diren eta espero zen sinplifikazioa egiten duten egiaztatu du.

Coster eta Kauchak-ek (2011a) BLUEz gain, testuen trinkotasuna neurtzeko erabili diren beste bi neurri erabiltzen ditu: *Simple String Accuracy* (SSA) neurria eta hitzen gainean kalkulatutako F neurria. Barlacchi eta Tonelli-k (2013) TER erabiltzen dute, TER-Plus tresnarekin. Bach *et al.*ek (2011), berriz, itzulpen eta laburpen automatikoetan erabiltzen diren AveF10, ROUGE-2 eta ROUGE-4 metrikekin batera Flesch-Kincaid konplexutasun-neurria erabili dute.

IAko metriken eta erabiltzaileen ebaluazioen korrelazioak ere aztertuak izan dira. Štajner *et al.*ek (2014b) METEOR, T-BLEU, eta TINE metriketan gramatikaltasuna korrelatuta dagoela esan dute. Esanahiari eusteari dagokionez, kosinua, METEOR, T-BLUE, SRL eta TINE ondo korrelatuta daudela aipatu dute. Lan horretan bertan, esaldiak sailkatzeko taxonomia bat proposatu dute: 1) onargarriak direnak, 2) post-edizioa behar dutenak eta 3) baztertu behar direnak.

Hala ere, neurri horiek egokiak ote diren komunitatean zalantzakotzat jotzen da, eta, adibidez, Štajner *et al.*ek (2015a) diote BLUE ez dela baliagarria esaldien antzekotasuna ebaluatzeko, eta BLUE metrika hobetzen saiatzeak esaldien jariotasuna okertzea eta esanahia aldatzea ekar dezakeela. Orobat, sistemaren entrenamendurako datu-kopuruak sistemaren performatzian eraginik ez duela adierazi dute.

2.5.4 Bestelako metodoak

Beste tresnetan sinplifikazioak duen eragina neurtu izan da, adibidez, hinditik ingeleserako itzulpen automatikoan (Mishra *et al.*, 2014). Sinplifikatutako esaldiak eta itzulitakoak ebaluatzeko, IAko neurriak eta hizkuntzalarien ebaluazioak erabili dituzte.

Sistemak ebaluatzeko beste metodo bat neurketak corpusaren kontra egitea da. Ebaluazio-mota horrek eskatzen duen baliabidea da eskuzko corpus sinplifikatu bat izatea urre-patroi edo *gold standard* bezala erabiltzeko. Modu honetan sinplifikazio-eragiketak eta sinplifikazio-erregelak ondo aplikatzen diren aztertu ohi da.

Candido *et al.*ek (2009) eskuz sinplifikatutako corpus baten kontra eragiketa guztiak banan-banan ebaluatu dituzte, doitasuna, estaldura eta F neurria erabiliz. Horretaz gain, esaldiak zuzen sinplifikatu diren ebaluatzeko, eskuzko ebaluazioaren beharra aurreikusi dute esaldiak benetan sinplifikatu diren jakiteko, Aluísio *et al.*en (2008b) aipatu bezala. Sistema ebaluatzeko, 143 esaldiz osatutako erreferentzia-corpusa eraiki dute eta bertan interesgarriak diren egitura sintaktikoak jaso dituzte Gasperin *et al.*ek (2010). Corpuseko esaldiak eskuz sinplifikatu dira sinplifikazio-erregelak jarraituz. Perpaus adberbialen sinplifikazio-erregelak ebaluatzeko, erreferentzia corpuseko esaldi sinplifikatuen eta jatorrizko esaldien konparazioa bi etiketatzaileek egin dute, eta horiei hiru etiketa jartzea eskatu diete: 0) esaldi sinplifikatuaren esanahia aldatzen da 1) esaldi sinplifikatuaren esanahia ez da aldatzen, baina ez da irakurtzeko errazagoa 2) esaldi sinplifikatuaren esanahia ez da aldatzen eta irakurtzeko errazagoa da. Sinplifikatutako esaldiak ebaluatzeko, berriz, *Levenshtein* (edizio) distantzia erabiliz erkatu dituzte erreferentzia corpuseko esaldi sinplifikatuekin. Erregelen aplikazio-hurrenkera ere ebaluatu dute *Levenshtein* distantzia erabiliz.

Sinplifikazio lexikala ebaluatzeko, Cohen-en *kappa* indizean oinarritutako etiketatzaileen arteko adostasun neurria proposatu dute Specia *et al.*ek (2012). Horrela, etiketatzaileen arteko adostasuna kontrastatu eta sistemen arteko konparazioa egin dute urre-patroiaren kontra.

Erregelak ere ebaluatu dira. Bott *et al.*ek (2012c), erregelak zein kasutan aplikatu diren kontuan hartuz, erregela testuinguru egokian aplikatu den eta emaitza ona izan duen ebaluatu dute. Ondoren, doitasuna, F neurria eta erregela zenbat aldiz aplikatu den kalkulatu dute, etiketatu dituzten 262 esaldietan. Evans *et al.*ek (2014), berriz, beraien erregelak hizkuntzalariek egindako erregelekin erkatu dituzte.

Ebaluaziorako neurriak eta metodoak bateratzeko asmoarekin, eta goian aipatutako esperimentuetan oinarrituta, Temnikova eta Maneva-k (2013) C-neurria *(Comprehension Score, C-score)* proposatu dute. C-neurria testuz testu kalkulatzen da eta hiru formula ditu: sinplea, osoa eta testu tamainaren araberakoa. Formula bakoitzak aldagai batzuk hartzen ditu kontuan, eta testuen tamaina ezberdinen arabera aplika daiteke bata edo bestea.

Aipatutako lanetan ikusi izan dugun bezala, oro har metodo bat baino gehiago erabiltzen da sistemen errendimendua eta irteera ebaluatzeko. Gehien erabiltzen diren ebaluazio-motak erabiltzaile bidezkoa eta IAko neurriak badira ere, moduluz moduluko ebaluazioa, corpusaren kontrako ebaluazioa, testuen konplexutasunaren analisia eta beste testuetan duten eragina ikustea dira beste ebaluazio posible batzuk.

2.6 Laburpena

Kapitulu honetan, testuen sinplifikazioa zer den ikusi dugu eta TSAn murgildu gara. HPko ikerketa-lerro honetan egin diren lanak ezagutu ditugu, helburu-taldearen, hizkuntzaren eta teknikaren edo metodoaren arabera sailkatuz. TSA egiteko beharrezkoak diren baliabideak eta corpusak ere aztertu ditugu, eta sistemak eta euren ebaluazioak azaldu ditugu. Horretaz gain, testuen konplexutasuna analizatzen dituzten lanak ere aurkeztu ditugu.

Ikusi dugu sintaxia sinplifikatzen duten eta ingeleserako ez diren sistemak oro har erregeletan oinarrituta daudela, hasierako lanak salbu. Hala ere, azken urteetako joera da sintaxia erregelen bidez sinplifikatzea, eta lexikoa sinplifikatzeko, berriz, teknika estatistikoak erabiltzea. Izan ere, mota honetakoak dira azken urteotan proposatzen ari diren sistemak. Sistemek sortutako esaldi sinplifikatuen artean egokiena aukeratzeko ere, teknika estatistikoak erabili dira.

Sistemak ebaluatzeko, berriz, hainbat ebaluazio-teknika erabili diren arren, egileek, oro har, erabiltzaileen bidezkoa, IAko eta konplexutasunaren analisiaren bitarteko neurrien bidezkoak gailendu dira, beste tresnetan sinplifikazioak duen eragina ere neurtu den arren. Erabiltzaileen bidezko ebaluazioan, oro har, galdeketetan kontuan izan diren parametroak sinpletasuna, gramatikaltasuna eta jatorrizkoaren esanahia gordetzea izan dira.

Beraz, egun dauzkagun lanetan oinarrituta, esan genezake ereduzko sis-

temak testua analizatuko lukeela, testuaren konplexutasuna ebaluatuko lukeela, sintaxia erregelen bitartez sinplifikatuko lukeela eta lexikoa estatistika bidez. Sor ditzakeen esaldi sinplifikatuetatik onenak aukeratzeko, teknika estatistikoak erabiliko lituzke. Sistema hori galdeketa bidez eta beste tresna batean ebaluatu beharko litzateke.

Behin atzerriko hizkuntzetan zein lan egin diren jakinik, 3. kapituluan euskarazko egitura konplexuen azterketa aurkeztuko dugu.

3

Euskarazko egitura sintaktiko konplexuen azterketa linguistikoa

Kapitulu honetan tesi-lan honetan sintaxian egitura konplexutzat hartu ditugun fenomenoak aurkeztuko ditugu. Lehenik, gure ustez, konplexuak zeintzuk diren adieraziko dugu eta, ondoren, egitura horiek sinplifikatzeko proposamenak azalduko ditugu. Proposamen horiek formulatzeko, corpusen azterketak egin ditugu eta azterketa horietatik ikasitakoak izango ditugu kapitulu honetan hizpide.

3.1 Sarrera

Testuen konplexutasuna aztertzeko erabili diren irizpideen artean esaldien luzera, hitzen silaba-kopurua, esaldiko perpaus-kopurua, eta abar daude. Lan honetan, zehazki honakoak hartu ditugu konplexutzat: perpaus elkartu mota guztiak (aditz bat baino gehiago dituzten perpausak), aposizio-sintagmak eta egitura parentetikoak. Erabaki horren arrazoia da egitura[1] horiek esaldiak luzeagoak egiten dituztela eta esaldiko ideia bat baino gehiago adierazten dutela[2]. Esaldi luzeek HPko tresnetan arazoak sortzen dituzte eta hizkuntzak ikasten ari diren pertsonei ere zailak egiten zaizkie. Hori dela-eta, esaldi sin-

[1] *Egitura* hitza erabiliko dugu aztergai dugun fenomeno bakoitza izendatzeko, hau da, fenomeno bakoitzaren errealizazioa ezberdintzeko. Adibidez, *-en ondoan* eta *-tu ondoan* ondokotasuna adierazten duten denbora-perpaus adberbialen egiturak dira.

[2] Irakurketa errazeko gidalerroetan ere esaldiko ideia bat ematea gomendatzen da (Freyhoff *et al.*, 1998; Mencap, 2000; Action eta Network, 2011).

pleak (perpaus bakunak) sortzea izango dugu helburu. Esaldi sinple horiek jatorrizko esaldiaren esanahiari eutsi behar diote eta jatorrizkoaren informazioa ahalik eta gutxien galdu behar dute.

Konplexutasunaren azalpena osatzeko eta euskarazko sinplifikazio-proposamenak egiteko, ingeleseko (Siddharthan, 2002) eta Brasilgo portugeseko (Specia *et al.*, 2008; Aluísio *et al.*, 2008a, b) lanetan oinarritu gara. Lan horietan sinplifikatu dituzten fenomeno berak aztertu ditugu euskaraz; fenomenoei euskal ordaina emanda, ingelesez aztertu dituztenak dira: perpaus erlatiboak, perpaus adberbialak (menderagailurik gabe), perpaus koordinatuak, mendeko perpausak (menderagailuarekin), perpaus korrelatuak, partizipio-sintagmak, aposizioak eta ahotsa; eta portuguesez: aposizioak, perpaus erlatiboak, mendeko perpausak, perpaus koordinatuak, aditz ez-jokatuak dituzten esaldiak eta ahots pasiboa. Fenomeno horiek euskarazko honakoekin bat datoz: perpaus koordinatuekin, mendeko perpausekin (osagarriak, erlatiboak eta adberbialak; jokatuak zein ez-jokatuak) eta aposizio-sintagmekin.

Hurrengo ataletan konplexutzat hartu ditugun euskarazko fenomenoen egituren sinplifikazio-proposamenak aurkeztuko ditugu, baina lehenik azterketa hori egiteko baliabideak eta definitutako metodologia ere deskribatuko ditugu.

3.2 Azterketa linguistikoa egiteko baliabideak eta metodologia

Euskarazko egitura konplexuen sinplifikazio-proposamenak egiteko, *Consumer* corpusa, *Euskararen Prozesamendurako Erreferentzia Corpusa* (EPEC), *Wikipedia* entziklopedia, *Elhuyar (T-comp)* corpusa eta *Lexikoaren Behatokia* corpusa erabili ditugu.

Consumer corpusa Eroskik argitaratzen duen *Consumer* aldizkariko testuekin osatuta dago (Alcázar, 2005). 1998 eta 2009 bitarteko 131 alek osatzen dute eta horietan 2.590 artikulu biltzen dira. Corpus espezializatua da, kontsumo arlokoa hain zuzen, eta bere berezitasuna da lau hizkuntzetako (gaztelania, euskara, galiziera eta katalana) testuak biltzen dituela. Hizkuntza horien arteko esaldiak automatikoki lerrokatu dira. Berezitasun hori dela-eta, itzulpen automatikoko lanetan erabili da corpusa (Labaka, 2010). Itzulpen automatikoko esperimentu horietan arazo gehien eman dituzten esaldiekin eta esaldi luzeekin osatutako lagina erabili dugu sinplifikazio-erregelak pro-

posatzeko. Lagin horrek 196 esaldi ditu eta esaldi horietatik 173 esaldik mendeko perpausak dituzte. Esaldi luzeenak 63 hitz ditu, eta laburrenak, berriz, 22 hitz. Guztira 7.759 hitz daude.

EPEC corpusa (Aduriz *et al.*, 2006a) euskara batuan idatzitako 300.000 hitzez osaturiko testu-bilduma da. Testu horiek *XX. mendeko euskararen corpus estatistiko*tik[3] eta *Euskaldunon Egunkaria*tik[4] hartu dira. Maila desberdineko informazio linguistikoarekin etiketatu da eskuz eta automatikoki: morfologia (Aldezabal *et al.*, 2007b), sintaxia (Aldezabal *et al.*, 2007a), semantika (Agirre *et al.*, 2006; Aldezabal *et al.*, 2010) eta pragmatika (Iruskieta *et al.*, 2011). Corpus hori euskararen prozesamendurako erreferentzia-corpusa da, alegia, euskararen prozesamendurako hainbat tresna garatzeko eta ebaluatzeko erabiltzen da.

Tesi-lan honetan, EPEC sinplifikazio-erregelak proposatzeko eta corpus azterketa kuantitatiboa egiteko erabili dugu. Zehazki, EPECen bi lagin erabili ditugu: i) EPEC-DEP zuhaitz-bankuko esaldiak, mendeko perpausak eta perpaus koordinatuak dituztenak, eta ii) EPEC corpusa etiketatzean luzeak zirelako eskuz etiketatu gabe geratu ziren esaldiak (esaldi luzeak). EPEC-DEP zuhaitz-bankuak (Aranzabe, 2008) EPECeko 200.000 hitz hartzen ditu. Hitz horiek eskuz etiketatu dira Dependentzia Gramatikaren Teoria (Tesnière, 1959) jarraituz; etiketatze-lan horretan esaldiko hitzak binaka lotuz esaldi bakoitzaren zuhaitz sintaktikoa (dependentzia-zuhaitza ere esaten zaio) lortu da. Zuhaitz horietan dependentzia-etiketen bidez adabegietan dauden hitzen arteko gobernatzaile/mendeko erlazioak irudikatzen dira, eta bi hitzen arteko loturan mendekoak betetzen duen funtzio sintaktikoa adierazten da. EPECeko esaldi luzeen laginean, berriz, 595 esaldi dauzkagu; bertan mendeko perpaus bat gutxienez duten esaldiak 488 dira. Esaldirik luzeena 138 hitzekoa da, eta laburrena, berriz, 14 hitzekoa. Guztira 18.490 hitz daude.

Wikipedia entziklopedia librea sarean auzo-lanean edo elkarlanean egiten den baliabidea da. *Wikipedia* erabiltzearen arrazoia beste domeinuetan ager daitezkeen fenomeno konplexuak aztertzea izan da eta beste corpusetan oinarrituta proposatu ditugun erregelak domeinu entziklopedikoan ere baliagarriak diren ikustea. *Wikipedia*n aurkitu dugun fenomeno interesgarria egitura parentetikoena izan da eta horiek aztertzeko, artikuluen lehendabiziko paragrafoa izan dugu aztergai. Izan ere, *Wikipedia*ren gidalerroen arabera, bertan aurkitzen da artikuluaren informazio nagusia. Azterketa linguistikoa egiteko

[3] http://www.euskaracorpusa.net/ (2011n atzituta)
[4] http://www.egunero.info (2002an atzituta)

erabiltzeaz gain, Biografix (ikus 6. kapitulua) entrenatzeko eta ebaluatzeko ere erabili dugu.

Elhuyar (T-comp) corpusa *Elhuyar* aldizkaritik jasotako 200 testuk osatzen dute, horietatik 100 albisteak dira eta beste 100 erreportajeak (guztira 161.161 hitz). Testu horietan Zientzia eta Teknologiako dibulgazio-artikuluak jasotzen dira. Corpus hori ErreXail (ikus 5. kapitulua) entrenatzeko sortu bada ere, EPECen agertzen ez diren edo maiztasun gutxi duten egiturak aztertzeko ere erabili dugu. Zientzia eta Teknologia domeinuko testuetan erregelek nola jokatzen duten ikusteko eta Euskarazko Testu Sinplifikatuen Corpusa (ikus 7. kapitulua) sortzeko abiapuntutzat ere hartu dugu, corpus honetatik atera baititugu jatorrizko testuak.

Lexikoaren Behatokia corpusa[5] Euskaltzaindiak lexikoa monitorizatzeko erabiltzen duen corpusa da. 2014. urte bukaeran 41.773.391 testu-hitz zeuzkan eta bertsio horretatik sortu dugu lemen[6] maiztasun-zerrenda. Lemen maiztasun-zerrenda, esaldiak sinplifikatzean ezabatutako elementuak berreskuratzeko definitu ditugun txertatze-elementuak aukeratzeko erabili dugu. Zerrenda horrek corpusean gehien eta gutxien erabili diren lemak ematen dizkigunez, sinplifikazio lexikala egiteko baliagarria izango da.

Azterketa linguistikoan erabili dugun beste baliabide bat gramatika izan da. Bi gramatika izan dira batez ere lan honen oinarri. Lehenengoa, Euskaltzaindiaren *Euskal Gramatika Lehen Urratsak* (EGLU) gramatika izan da, eta, batez ere, mendeko perpausei dagozkien liburukiak (Euskaltzaindia, 1999, 2005, 2011) erabili ditugu. Bigarrena, berriz, *Sareko Euskal Gramatika* (SEG)[7] izan da, zeinetatik mendeko perpausei dagozkien ataletako informazioa erabili dugun gehienbat. Kontsulta moduan, *A Brief Grammar of Euskara, the Basque Language* gramatikaren sareko bertsioa[8] (Laka, 1996), *A Grammar of Basque* (Hualde eta Ortiz de Urbina, 2003) eta *Euskal Gramatika Laburra* (EGLA) (Euskaltzaindia, 2002) ere erabili ditugu.

Baliabide horietan oinarrituta, eskuzko azterketa egiteko erabilitako metodologia honako hau izan da:

1. Aztergai izan dugun fenomenoari dagozkion esaldiak erauzi. Uneko fe-

[5] http://lexikoarenbehatokia.euskaltzaindia.net/aurkezpena.htm (2015eko apirilean atzituta)

[6] Lema hitzari esanahia ematen dion morfema da, hitza osatzeko ezinbestekoa dena eta lexikoetan eta hiztegietan agertzen dena. Adibidez, *etxearen* hitzaren lema *etxe* da.

[7] http://www.ehu.eus/seg/aurkezpena (2015eko apirilean atzituta)

[8] http://www.ehu.eus/eu/web/eins/a-brief-grammar-of-euskara (2015eko apirilean atzituta)

nomenoaren esaldi-bildumak osatu ditugu, fenomeno horiei soilik erreparatzeko.

2. Esaldi horiek eskuz aztertu. Horretarako, hainbat galdera egin ditugu: Menderagailu ezberdinak erabiltzen dira? Erlazio[9] ezberdinak al daude? Gramatiketan agertzen diren egitura guztien adibideak dauzkagu?

3. Beharrezkoa izanez gero, azpisailkapenak egin. Gramatiketan agertzen den informazioa kontuan izan dugu ataza horretan.

4. Sinplifikazio-proposamenak egin. Beharrezkoa izan denean, beste hizkuntzetan egin diren proposamenak kontuan izanda, multzoari oro har ondo datorkion proposamena egin dugu.

5. Sinplifikazio-proposamenak beste domeinu batzuetako corpusetan edo/eta beste tresnaren batean (itzultzaile automatikoan edo galdera-sortzailean) baliagarriak diren ikusi. Horrela, egindako proposamena domeinu bakar batera mugatzen ez dela ziurtatu dugu.

6. Sinplifikazio-proposamenak dokumentatu. Fenomeno bakoitzari dagozkion sinplifikazio-proposamenak jaso ditugu, horiekin batera arazo posibleak aipatuz.

Eskuzko azterketa horretaz gain, EPEC-DEP corpusean mendeko perpaus ezberdinen kopuruak[10] (azterketa kuantititatiboa) atera ditugu (ikus 3.1 taula). Kopuru horiek fenomeno bakoitzaren erabilera eta banaketa erakusten dute: mendeko perpaus-motarik erabilienak perpaus adberbialak dira, % 38,71ko maiztasunarekin. Perpaus osagarriek eta erlatiboek corpusaren % 34,63 eta % 25,23 osatzen dute hurrenez hurren. Bestelako perpausek % 1,43 osatzen dute.

Fenomenoen emaitza xehatuak dagozkien ataletan emango ditugu.

3.3 Sinplifikazio-proposamenak

Atal honetan euskarazko egitura konplexuak sinplifikatzeko proposamenak aurkeztuko ditugu. Lehenik, perpaus elkartuak izango ditugu aztergai, hau

[9] Erlazioak emendioa, aurkaritza, denbora, kausa eta abar dira.

[10] Perpaus koordinatuen, aposizio-sintagmen eta egitura parentetikoen emaitzak ez ditugu 3.1 taulan eman, fenomeno horien erlazioak corpusean unibokoak ez direlako. Izan ere, fenomeno horiek etiketatzeko konplexuak dira.

Fenomenoa	Ehunekoa
Perpaus erlatiboak	25,23
Perpaus adberbialak	38,71
Perpaus osagarriak	34,63
Bestelakoak	1,43

3.1 taula – Mendeko perpausen banaketa corpusean

da, perpaus koordinatuak eta mendeko perpausak. Mendeko perpausen artean perpaus osagarriak (osagarri funtzioa dutenak, hau da, subjektu, objektu eta zehar-objektu funtzioak), perpaus erlatiboak (izenlagun funtzioa dutenak) eta perpaus adberbialak (adizlagun funtzioa dutenak) aztertuko ditugu. Perpaus osagarrien azpiatalean, horien antzera sinplifikatzen diren bestelako egitura bat eta usteak edo adierazpenak adierazten dituzten postposizio-sintagmak ere jaso ditugu. Mendeko perpausak aztertu ondoren, modifikatzaileak diren aposizio-sintagmak eta egitura parentetikoak azalduko ditugu. Perpaus adberbialen atalean, EPEC-DEP corpusean egin dugun corpus-azterketaren emaitzak ere emango ditugu.

Sinplifikazio-proposamenekin, aurretik aipatu bezala, aditz bakarra duten esaldiak lortu nahi ditugu. Esaldi horiek, gainera, eutsi behar diote ahal den bezainbeste jatorrizko esaldiaren esanahiari. Oro har, helburu hori lortzeko sinplifikazio-proposamen horiek urrats hauek jarraitzen dituzte:

1. Esaldian dauden perpausak, landutako postposizio-sintagmak, aposizio-sintagmak edo egitura parentetikoak banatu. Hurrengo urratsean landuko ditugun unitateak (perpausak, sintagmak edo tartekiak) lortuko ditugu.

2. Erlazio-markak edo juntagailuak ezabatu. Erlazio-markak menderagailuak, kasu-markak, postposizioak eta elementu askeak dira; adibidez, *-enetik, -en ostean, ba- ere* eta abar. Oro har, mendeko perpausen erlazio-markak mendeko perpausetatik ezabatuko ditugu.

3. Ezabatutako erlazio horiek berreskuratuko dituzten elementuak txertatu. Elementu horiek *txertatze-elementuak* (TE) dira, eta, oro har, perpaus nagusietan txertatuko ditugu. TEak aukeratzeko, esanahiaz gain, bere lemaren maiztasunak ere aintzat izango ditugu. Izan ere, ez dugu nahi adierazten den erlazioa zailagoa egin TE desegokia aukeratu dugulako.

Perpaus adberbialen sinplifikazio-proposamenetan, txertatze-elementu alternatiboak ere definitu ditugu, testu batean erlazio bat behin baino gehiagotan agertzen bada, sinplifikatzean testuan behin eta berriro hitz bera[11] errepikatzeak sor dezakeen monotoniarekin hausteko. TE alternatiboak erabiliko ditugu baldin eta jada esaldian TE lehenetsia (maiztasun altuena duena) aurkitu badugu edo testuan jada erabili badugu. Aditz ez-jokatuak dituzten perpausetan, horietaz gain, aditz ez-jokatua aditz jokatu bihurtuko dugu (hau da, aspektua eta aditz laguntzailea txertatu). Baina ataza hori ez da batere erraza; izan ere, aditzaren argumentu inplizituak, pertsonak, denbora eta modua zein diren jakin behar dugu. Informazio hori aditzetik bertatik (argumentuak zein kasutan jarri), mendeko esalditik (zein pertsona dauden) eta perpaus nagusiaren aditzetik (aldia eta aspektua) lortu behar dugu.

4. Esaldi berriak ordenatu. Aurreko urratsean sortutako esaldi sinplifikatuak testuan ordenatuko ditugu. Horretarako, perpaus adberbialen kasuan, corpus-azterketa kuantitatiboan oinarrituko gara.

5. Esaldiak zuzendu (ortografia eta ortotipografia). Esaldi guztiak sinplifikatu ondoren, egon daitezkeen ortografia-, estandarizazio- eta gramatika-akatsak zuzenduko ditugu; izan ere, sortu nahi ditugun esaldiak ulermenean eta tresnen prozesamenduan eraginkorrak izan daitezen, zuzenak izan behar dira. Esaldi guztien amaieran puntua jarriko dugu eta hurrengoa hizki larriz hasiko dugu, kontrakorik azaldu ezean. Perpaus adberbialetan, mendeko perpausaren ondoren koma badago, TEaren ondoren ere koma jarriko dugu; TEak lokailuak badira, beti txertatuko dugu koma.

Egindako proposamenean jatorrizko esaldietan egitura-aldaketak egingo ditugu. Hau da, jatorrizko perpausean legokeen mendekotasuna (eta horien sakonera) kenduko dugu esaldi laburragoak lortzeko. Urrats horien ondorioz, jatorrizko esaldiaren zuhaitz sintaktikoa aldatu egingo da. Beraz, sinplifikazio-mota hori sinplifikazio sintaktikoa da.

Perpaus adberbialen sinplifikazio-proposamenetan, aipatutako sinplifikazio sintaktikoaren proposamenaz gain, beste sinplifikazio-mota bat aurkeztuko dugu: ordezkapen sintaktikoen sinplifikazioa. Ordezkapen sintaktikoen

[11] Irakurketa errazeko gidalerroek kontzeptu bera adierazteko beti hitz bera erabili behar dela proposatzen dute, baina guk txertatze-elementu alternatiboak proposatzen ditugu testuak irakurlearen mailara egokitu daitezkeelako.

sinplifikazioan maiztasun gutxiko egiturak baliokide diren maiztasun handiagoekin ordezkatuko ditugu. Sinplifikazio lexikalaren urratsak (sininomoekin ordezkatu maiztasunak erabiliz) egin arren, sintaxian gauzatuko da eta ez da esaldiaren zuhaitz sintaktikoan egitura-aldaketarik egingo. Ondorioz, esaldiaren zuhaitz sintaktikoaren sakonera mantenduko da. Maiztasunetan oinarritzeko, corpus-azterketa kuantitatiboaren emaitzak erabiliko ditugu eta aditz ez-jokatua duten perpausentzat soilik proposatuko ditugu.

Hurrengo azpiataletan aztergai izango ditugun egituren adibideak EPEC corpusekoak dira, kontrakorik esan ezean. Adibide horietan, egitura letra lodiagoz nabarmendu dugu eta txertatu ditugun elementuak letra etzanez. Adibideen (a) multzoak jatorrizko esaldia adierazten du eta (b) multzokoak gure sinplifikazio-proposamenen arabera sinplifikatu ditugun esaldiak dira. Adibideetan eta azalpenetan erabili ditugun laburtzapenak eta dagokien esanahia 3.2 taulan jaso ditugu.

Mota	**Laburtzapena**	**Esanahia**
Egitura-mota	p_i	perpausa
	s_i	sintagma (aposizio-, izen- nahiz postposizio-sintagmak)
	t_i	tartekia
Hurrenkera-mota	koord$_{jat1}$-koord$_{jat2}$	lehenik, lehen perpaus koordinatua eta ondoren, bigarren perpaus koordinatua (jatorrizko hurrenkera)
	nagusia$_{jat}$-mendekoa$_{jat}$	lehenik, jatorrizko esaldian perpaus nagusia zena eta ondoren, jatorrizko esaldian mendeko perpausa zena
	mendekoa$_{jat}$-nagusia$_{jat}$	lehenik, jatorrizko esaldian mendeko perpausa zena eta ondoren, jatorrizko esaldian perpaus nagusia zena
	nagusia$_{jat}$-apos$_{jat}$	jatorrizko esaldian perpaus nagusia zena eta ondoren, jatorrizko esaldian aposizio-sintagma osatzen zuten izen-sintagmekin osatutakoa
	post$_{jat}$-nagusia$_{jat}$	lehenik, jatorrizko esaldian postposizio-sintagma zenetik sortutakoa eta ondoren, jatorrizko esaldian perpaus nagusia zena
Txertatze-elementuen kokapena	txertatze-elementua$_{nag}$	jatorrizko esaldian perpaus nagusia zen esaldian txertatu behar den TEa edo TE alternatiboa

(Jarraipena hurrengo orrialdean)

Mota	Laburtzapena	Esanahia
	txertatze-elementua$_{men}$	jatorrizko esaldian mendeko perpausa zen esaldian txertatu behar den TE edo TE alternatiboa

3.2 taula – Adibideetan erabilitako laburtzapenen eta ikurren esanahia

3.3.1 Perpaus koordinatuak

Koordinazioan edo juntaduran maila bereko osagaiak (sintagmak zein perpausak) elkartzen dira eta juntagailuekin edo juntagailurik gabe (alborakuntza bidez) elkar daitezke. Azpiatal honetan, aurreko atalean deskribatutako baliabidetan aurkitu ditugun mota honetako perpausak izango ditugu aztergai.

Emendiozko perpaus koordinatuak *eta* juntagailuaren bidez lortzen dira eta osagaiak gehitzea dute helburu. Perpaus horiek sinplifikatzeko proposamena (1) adibidean erakutsiko dugu. Urrats hauek egin ditugu:

1. Perpausak banatu: [Erakunde horren arabera, EAEko langabezi tasa 1999ko lehen hiruhilekoan %15,5koa izan zen]$_{p_1}$ [eta 2000ko epe berean %14,8ra jaitsi da.]$_{p_2}$
2. p$_2$ perpausetik juntagailua ezabatu: eta -> ø
3. Esaldi berriak ordenatu: koord$_{jat1}$-koord$_{jat2}$
4. Puntuazio-markak egokitu eta zuzendu

(1) a. Erakunde horren arabera, EAEko langabezi tasa 1999ko lehen hiruhilekoan %15,5koa izan zen **eta** 2000ko epe berean %14,8ra jaitsi da.

b. i. Erakunde horren arabera, EAEko langabezi tasa 1999ko lehen hiruhilekoan % 15,5ekoa izan zen.

ii. 2000ko epe berean % 14,8ra jaitsi da.

Alborakuntzako (juntagailurik gabeko juntadura) perpaus koordinatuak ere horrela sinplifikatu ditugu.

Aurkaritzako juntadura adierazteko, *baina* eta *baizik* juntagailuak erabiltzen dira eta perpauseko osagaiak aurka edo kontra jartzea dute helburu.

Aurkaritzako juntaduraren sinplifikazio-proposamena (2), emendiozko juntaduran ez bezala, ezabatutako juntagailua berreskuratu dugu eta bigarren esaldiaren hasieran txertatu dugu esaldiaren esanahiari eusteko.

(2) a. Irlandako Poliziak RIRAko 14 ustezko kide atxilotu ditu azken astean, **baina** horietako zazpi jada aske utzi ditu, kargurik gabe.
b. i. Irlandako Poliziak RIRAko 14 ustezko kide atxilotu ditu azken astean.
ii. *Baina* horietako zazpi jada aske utzi ditu, kargurik gabe.

Hautakaritzako juntaduran *edo*, *ala*, *nahiz* eta *zein* juntagailuak dituzten adibideak ageri dira. Juntagailu horiek, oro har, osagai baten edo bestearen artean aukera egitera behartzen gaituzte. Corpusean aurkitutako adibide gehienak *edo* eta *ala* juntagailuekin osatzen dira. Beraz, aurkaritzakoekin egin dugun bezala, jatorrizko perpausek daukaten juntagailua bigarren koordinatuaren hasieran txertatu dugu.

Era berean, perpaus koordinatuetan ohikoa da bigarren perpausaren aditza elidituta egotea, batez ere aditz hori lehenengoan agertu den bera bada. Kasu horietan bigarren perpausean lehenengo perpauseko aditza berriro txertatu dugu, alegia kopiatu dugu. Proposamen hori (3) adibidean ikus dezakegu.

(3) a. Tropelean gogor aritu ziren esprinterren taldeak gasteiztarra harrapatzeko, baina atzo ez ø.
b. i. Tropelean gogor aritu ziren esprinterren taldeak gasteiztarra harrapatzeko.
ii. Baina atzo ez *ziren gogor aritu*.

Beraz, (3) adibideko bigarren perpaus koordinatuan *gogor aritu* aditza txertatu dugu perpausean elidítutako aditza esplizitu egiteko. Adibide horretan ez dugu aditz laguntzailearen komunztadura aldatu, baina ezezko perpausa denez, barne-hurrenkera aldatu dugu.

3.3.2 Perpaus osagarriak

Perpaus osagarria menderakuntza bidez sortzen den perpaus-mota bat da. Perpaus osagarriak perpaus konpletiboek eta zehar-galderek osatzen dituzte eta perpaus nagusiaren argumentuak dira. EPEC-DEPeko mendeko perpausen % 34,63 osatzen dute; horietatik % 94,34 konpletiboak dira eta % 5,66

zehar-galderak. Azpital honetan perpaus osagarriez gain, perpaus osagarriak bezala sinplifikatuko diren bestelako egiturak (*-enez + aditz diskurtsiboak*) eta postposizio-sintagmak (*-en arabera, -en hitzetan,-en adierazpenetan*) ere azalduko ditugu. Hain zuzen, egiturok perpaus osagarrietan adierazten den bera (objektua) iragartzen da perpaus nagusi forman.

Perpaus konpletibo jokatuek dituzten ohiko menderagailuak dira *-ela* baiezko perpausetan eta *-enik* ezezko perpausetan. Perpaus horiek (4) sinplifikatzeko proposamena estilo-aldaketa da, zehar-estilotik estilo zuzenerako aldaketa, hain zuzen ere. Hauek dira jarraitu ditugun urratsak:

1. Perpausak banatu: [Eri, gaixorik naizela]$_{p_1}$ [esan genezake...]$_{p_2}$
2. Mendeko perpausean erlazio-markak ezabatu: naizela -> naiz
3. Perpaus nagusian aditzaren aurretik "honako hau" txertatu
4. Esaldi berriak ordenatu: nagusia$_{jat}$-mendekoa$_{jat}$
5. Puntuazio-markak egokitu: nagusia$_{jat}$ esaldiaren amaieran bi puntu ipini (4b-i) eta mendekoa$_{jat}$ esaldia komatxoen artean jarri (4b-i)
6. Aditzaren pertsonak eta izenordainak aldatu, beharrezkoa bada

(4) a. Eri, gaixorik naiz**ela** esan genezake...
b. i. *Honako hau* esan genezake:
ii. "Eri, gaixorik naiz."

Baiezko perpaus konpletiboetan ez bezala, ezezkoetan (5) "honako hau" aditz laguntzailearen eta aditz nagusiaren artean txertatu dugu.

(5) a. Ez zait iruditzen agurra behar bezalakoa izan d**enik**.
b. i. Ez zait *honako hau* iruditzen:
ii. "Agurra behar bezalakoa izan da."

Aipatu nahi dugu erabili ditugun adibideetako esaldiak nahiko laburrak direla eta horiek sinplifikatzeko proposamenek testuaren erritmoa moteldu dezaketela, baina adibide horiek erabili ditugu zaila baita perpaus konpletiboak soilik dituzten esaldiak corpusean aurkitzea. Orduan, horren ondorioz erabaki dugu ez dela esaldirik sinplifikatuko mendeko perpausa laburra bada. Beraz, kapitulu hasieran konplexutzat zer hartuko dugun esan dugu baina,

horretaz gain, kontuan hartu behar dugu esaldia sinplifikatu baino lehen, perpausek aditzaz gain gutxienez beste bi argumentu edo adjuntu izan behar dituztela. Murriztapen hori (luzera minimoa) perpaus-mota guztiei aplikatuko diegu. Hortaz, (4) eta (5) adibideak ez ditugu sinplifikatuko.

Zehar-galderek bere azpian galdera bat gordetzen dute. Perpaus jokatua duten zehar-galderen ohiko menderagailua *-en* da eta maiz *ea* partikula ere aurkitzen dugu. Galdetzaileak ere aurki daitezke. Aditz jokatuak dituzten zehar-galderak sinplifikatzeko, perpaus konpletiboekin egin dugun bezala, estilo zuzenera pasa ditugu. Hau da, zehar-galdera izatetik galdera zuzena egitea da egin dugun eraldaketa, (6) adibidean ikus daitekeen bezala. Jatorrizkoan *ea* partikula egonez gero, ezabatu egingo dugu.

(6) a. Fiskalak galdetu zidan ea desobedientzia zibila eraikuntza nazionalerako egiten genu**en**.
    b. i. Fiskalak *honako hau* galdetu zidan:
    ii. "Desobedientzia zibila eraikuntza nazionalerako egiten zenuten?"

Galdetzaileak aurkitu ditugunetan (7), galdetzaileok esaldi berriaren hasieran jarri ditugu (7b-ii) eta ondoren aditza jarri dugu.

(7) a. Segur aski, irakurleak galdetuko dio bere buruari ea lan honetatik **zer** atera daitekeen.
    b. i. Segur aski, irakurleak *honako hau* galdetuko dio bere buruari:
    ii. "Zer atera daiteke lan honetatik?"

Aditz ez-jokatuak dituzten perpaus konpletiboei dagokienez, aurki ditugun adizkien formak *-tzen, -tzera, -tzeko, -tzea + atzizkia, -tzea, -tzerik* eta *-tu izana* dira. Horiek sinplifikatzeko (8) aditz jokatuek dituzten perpausekin bezala egin dugu, baina aditz ez-jokatuak jokatu bihurtzea zaila denez, aditza jokatu gabe utzi dugu (letra xehez hasi dugu, gainera). Zehar-objektu funtzioa duten perpaus ez-jokatuetan *honako hau* txertatu ordez *honako honi* txertatzea proposatu dugu komunztadura mantentzeko. Puntuazio-markei dagokienez, kasu honetan ez ditugu komatxoak bi puntuen ondoren ipini.

(8) a. Honek, ia hilabete atzeratuko luke zuen aldetik idatzia jaso**tzea**...
    b. i. Honek, ia hilabete *honako hau* atzeratuko luke:
    ii. zuen aldetik idatzia jasotzea...

Predikatibo funtzio sintaktikoa (9) badute, ez ditugu sinplifikatu. Corpusean aurkitutako adibideak oro har laburrak dira eta ekintzak baino sentimenduak eta egoerak adierazten dituzte.

(9) a. Askotan, biak nahas**turik** ere erabili dira.
b. Oraingo honetan Alberto Couso benetan harri**tuta** zegoen.

Corpusean ez ditugu zehar-galdera ez-jokatu asko aurkitu eta aurkitu ditugun adibideetan beste mendeko batzuekin batera agertu izan dira. Instantzia gehiago izan arte, perpausok ez sinplifikatzea erabaki dugu.

Perpaus osagarriak ez diren, baina perpaus osagarrien antzera sinplifikatu ditugun bestelako egiturei dagokienez (*-enez + aditz diskurtsiboak* eta usteak edo adierazpenak adierazten dituzten postposizio-sintagmak), *-enez* sasimoduzko egitura aditz diskurtsibo batekin aurkitzean (10), *-enez* erlazio-marka ezabatu dugun perpausean *honako hau* txertatu dugu. Alegia, perpaus osagarrietan *honako hau* perpaus nagusian txertatu den bezala, hauetan *-enez* erlazio-marka zeraman mendekoan txertatu dugu behin erlazio-marka hori ezabatuta. Esaldi berrien hurrenkera, kasu honetan eta osagarrietan ez bezala, mendekoa$_{jat}$-nagusia$_{jat}$ da, mendekoak adierazten baitu nork esan duen. Puntuazio-markei dagokienez, mendekoa$_{jat}$ perpausaren amaieran bi puntu ipini ditugu eta nagusia$_{jat}$ perpausa komatxo artean eman dugu.

(10) a. Cal Dooley parlamentario demokratak adierazi zu**enez**, botoak berriz kontatzeko agindua emango dutela espero dute Goreren alderdikoek.
b. i. Cal Dooley parlamentario demokratak *honako hau* adierazi zuen:
ii. "Botoak berriz kontatzeko agindua emango dutela espero dute Goreren alderdikoek."

Postposizio-sintagmadun egiturei dagokienez honakoak hartu ditugu aztergai: *-en arabera* (edo bere sinonimoak: arau, arauaz, arauka, ereduz, araberan eta eredura) (11), *-en hitzetan* eta *-en adierazpenetan*. Egiturok perpaus nagusian azaldutakoa noren ustetan esaten den adierazten dute. Azpimarratu nahi dugu proposamen hori aurrera eramateko genitiboan dagoen hitza biziduna edo bizidunez osatutako entitate bat (erakundea, tokia...) izan behar dela.

Postposizio-sintagma horiek sinplifikatzeko, urrats hauek egin ditugu:

1. Postposizio-sintagma gainotzeko esalditik banatu: [UPNko lehendakariaren arabera]$_{s_1}$ ["Allik bere egin nahi ditu UPNren lorpenak, baina ez du inor engainatuko".]$_{p_1}$
2. Erlazio-markak ezabatu: ren arabera -> ø
3. Genitiboa kendu diogun hitzari ergatiboa txertatu: UPNko lehendakaria -> UPNko lehendakariak
4. "honako hau dio(te)"$_{s_1}$ txertatu, subjektuarekin komunztadura mantenduz: UPNko lehendakariak honako hau dio
5. Esaldi berriak ordenatu: post$_{jat}$-nagusia$_{jat}$
6. Zuzendu eta puntuazio markak egokitu: post$_{jat}$ esaldiaren amaieran bi puntu ipini eta nagusia$_{jat}$ komatxo artean eman

(11) a. UPNko lehendakaria**ren arabera**, "Allik bere egin nahi ditu UPNren lorpenak, baina ez du inor engainatuko".
b. i. UPNko lehendakariak *honako hau* dio:
ii. "Allik bere egin nahi ditu UPNren lorpenak, baina ez du inor engainatuko."

-en hitzetan, -en adierazpenetan eta halako postposizio-sintagmak *-en arabera* bezala sinplifikatu ditugu. *-en adierazpenetan...* aurkituz gero, "honako hau adierazi du(te)" txertatu dugu eta *-en hitzetan* topatuz gero, "honako hau esan du(te)/ dio(te)". Gogoan izan behar da estilo-aldaketek pertsonaren eta izenordainen egokitzapenak eska ditzaketela.

3.3.3 Perpaus erlatiboak

Perpaus erlatiboak izenlagun funtzioa duten mendeko perpausak dira. Euskaraz, perpaus erlatibo jokatuak sortzeko dauden hainbat estrategiaren artean guk corpusetan bi aurkitu ditugu: *-(e)n* menderagailuarekin sortzen diren erlatibo arruntak eta "zein erlatiboak", *zein* edo *non* izenordainekin eta *bait-* edo *-en* menderagailuekin sorturikoak. Perpaus erlatibo ez-jokatuei dagokienez, corpusean aurkitu ditugunak *partizipioa + -ta/-ika/-i/+ko*, *aditz modalak + -ko* eta *partizipioa + gabe + -ko* egiturekin osatutakoak dira. EPEC-DEP corpusean mendeko perpausen % 25,23 osatzen dute. Atal honetan, bi horiek nola sinplifikatuko ditugun azalduko dugu.

Perpaus erlatibo jokatuak sinplifikatzeko proposamenak egiterakoan, kontuan izan dugu perpaus erlatiboek elementu bereizgarri bat dutela: aurrekaria. Aurrekaria da, izan ere, bi perpausek amonkomunean duten elementua eta biak lotzen dituena. Perpaus erlatibo arrunten sinplifikazio-proposamena nola gauzatu dugun (12) adibidean adierazi dugu:

1. Jatorrizko esaldian dauden bi perpausak banatu: [Konstituzioari eta Estatutuari eskaini zaizkien]$_{p_1}$ [ihardunaldietan Zuzenbide Zibila bazterrean geratzen da.]$_{p_2}$
2. Erlazio-marka (kasu honetan, mendeko perpauseko aditzaren *-en* menderagailua) ezabatu: zaizkien -> zaizkie
3. Aurrekaria identifikatu (ihardunaldietan) eta bi perpausetan txertatu:
   - Mendekoan, aditzak eskatzen duen kasuan txertatu, adibidean absolutiboan (ihardunaldiak)
   - Nagusian, jatorrizko perpausean duen kasua mantenduz, erakuslearekin txertatu (ihardunaldi horietan)
4. Esaldi berriak ordenatu: mendekoa$_{jat}$-nagusia$_{jat}$
5. Puntuazio-markak egokitu eta ortografia akatsak zuzendu: (ihardunaldietan -> jardunaldiak/jardunaldi)

(12) a. Konstituzioari eta Estatutuari eskaini zaizkie**n** ihardunaldietan Zuzenbide Zibila bazterrean geratzen da.

b. i. Konstituzioari eta Estatutuari *jardunaldiak* eskaini zaizkie.

ii. *Jardunaldi horietan* Zuzenbide Zibila bazterrean geratzen da.

Aurrekaria entitate bat baldin bada, edo entitatea aposizioan agertzen bada (13)[12], ez dugu aurrekariarekin batera perpaus nagusian erakuslerik txertatu.

(13) a. JOAN den igandeaz geroztik Alberto Fujimori Peruko presidentearen aurka altxamendu militar bat gidatzen ari de**n** Ollanta Moises Humala teniente koronelak ez du uste bakarrik dagoenik (...)

[12](13) adibidean fenomeno bat baino gehiago gertatzen diren arren, perpaus erlatiboaren sinplifikazioari erreparatu diogu hemen.

b. i. Joan den igandeaz geroztik Alberto Fujimori Peruko presidentearen aurka altxamendu militar bat gidatzen ari da *Ollanta Moises Humala teniente koronela*.

ii. *Ollanta Moises Humala teniente koronelak* ez du uste bakarrik dagoenik (...)

"Zein erlatiboak" sinplifikatzean (14) mendeko perpauseko *zein* izenordaina aurrekariarekin ordezkatu dugu (izenordaina ezabatu eta aurrekaria txertatu). Aurrekaria izen arrunta denez, aurrekariak eta erakusleak osatzen duten sintagmari *zein* izenordainak zuen kasua jarri diogu. Aurrekaria entitatea bada, aurrekariari *zein* izenordainak duen kasua jarriko diogu. Perpaus erlatibo horiek azalpenezko kutsua dutenez eta informazio gehigarria ematen dutenez, esaldi berrien hurrenkera nagusia$_{jat}$-mendekoa$_{jat}$ da.

(14) a. 1873ko urriaren 6ko dekretu gehigarri batek, **zeina** Errepublikako Gobernu-Presidente Emilio Castelar-ek eta Estatuko Ministro, Jose de Carvajal-ek izenpetu baitzuten, bi pentsio gehitzen zituen merituzko plaza banarentzat, arkitektura eta sakongrabatuko alorretan.

b. i. 1873ko urriaren 6ko *dekretu* gehigarri batek bi pentsio gehitzen zituen merituzko plaza banarentzat, arkitekturako eta sakongrabatuko alorretan.

ii. *Dekretu hori* Errepublikako Gobernu-Presidente Emilio Castelar-ek eta Estatuko Ministro Jose de Carvajal-ek izenpetu zuten.

Aditz ez-jokatua duten perpaus erlatiboen (15) sinplifikazio-proposamena aditz jokatua dutenen antzekoa da. Desberdintasuna da aditz ez-jokatutik aditz jokatua sortu behar dugula, eta, horretarako, aditzaren informazioa izatea ezinbestekoa da.

(15) a. Zuzendaritzak Donostian 1995.eko azaroaren 20an egin**dako** bileran familiari eta bereziki Pakitari doluminak agertzea erabaki zuen.

b. i. Zuzendaritzak Donostian 1995. urteko azaroaren 20an *bilera* egin zuen.

ii. *Bilera horretan* familiari eta bereziki Pakitari doluminak agertzea erabaki zuen.

3.3.4 Perpaus adberbialak

Perpaus adberbialak, hainbat erlazio (denbora, kausa, baldintza...) zehazten dituzten adjuntuak dira. Euskaraz, denbora-, kausa-, kontzesio-, modu-, helburu-, baldintza- eta konparazio- perpausak bereizten ditugu. EPEC-DEP corpusean perpaus adberbialek mendeko perpausen % 38,71 osatzen dute, eta mendeko perpausik erabilienak dira. Multzo zabala osatzen dutenez, lehenik, zein erabilera duten aztertu dugu (azterketa kuantitatiboa) eta, ondoren, datu horietan eta eskuzko azterketan oinarrituta, perpaus adberbialen sinplifikazio-proposamenak egin ditugu. Azterketa horietatik kanpo utzi ditugu konparazio-perpausak, horietan mendeko aditza elidituta dagoelako eta EPEC-DEP corpusean ez direlako sistematikoki etiketatu.

Perpaus adberbialek hainbat motatako erlazioak azaltzen dituztenez eta mota bakoitzaren barnean ere azpierlazioak aurki daitezkeenez, atal-hasieran aipatutako lau urratsak betetzen baditugu ere, erabakiak hartzeko informazio osagarria behar izan dugu. Informazio osagarri lortzeko motibazioa (16) adibidearekin azalduko dugu. (16a) esaldian denbora-perpaus bat dugu eta (16b) adibidean kontzesio-perpaus bat. Bi esaldi horietan perpausak banatzeko eta erlazio-markak ezabatzeko arazorik ez dugu. Baina bi kasu horietan erlazioa mantentzeko txertatuko dugun elementua ezin da bera izan. Esaldi sinplifikatuen hurrenkera definitzeko, hurrenkera logikoak (kausa-ondorio, baldintza-ondorio...) eta kronologikoak ere errespetatzeko joera oten den ere jakin behar dugu.

(16) a. Maradona eta Castro lagunak dira, 1987ko udan Habanan elkar ezagutu zute**netik**.

b. Ondo antolatutako Jokoak izan dira, transporte publikoarekin arazo batzuk izan dire**n arren**.

Beraz, perpaus adberbialen sinplifikazio-proposamenak egiteko, beste fenomenoekin egin ditugun azterketez gain, azterketa gehigarriak egin ditugu. Alde batetik, Lexikoaren Behatokia corpusetik ateratako lemen maiztasunzerrenda erabili dugu erlazioak berreskuratuko dituzten elementuak (txertatze-elementuak, TEak) aukeratzeko; beste aldetik, corpus-azterketa kuantitatiboa egin dugu zein egitura erabiltzen diren (ordezkapen sintaktikoen sinplifikazioa egiteko) eta zein hurrenkeratan gertatzen diren (esaldi sinplifikatuak ordenatzeko) jakiteko. Informazio horrekin sinplifikazio-proposamenak osatu eta sendotu ditugu.

Corpus-azterketa kuantitatiboa egiteko, alde batetik, EPEC-DEP corpuseko (Aranzabe, 2008) esaldi guztiak erauzi ditugu eta, beste aldetik, EGLU liburukietan (Euskaltzaindia, 1999, 2005, 2011) agertzen diren perpaus adberbialen egiturekin zerrenda bat (egitura-zerrenda, A eranskina) osatu dugu. Zerrenda horretan dauden egiturak corpusean bilatu ditugu hiru galdera hauei erantzuteko:

1. Zerrendako zenbat egitura agertzen dira corpusean? (Agerpena)
2. Zenbatetan agertzen dira egitura horiek? (Maiztasuna)
3. Non agertzen dira perpaus nagusia kontuan hartuta? (Kokapena)

Lehenengo galderari erantzunez, perpaus adberbialak oro har deskribatzen dituzten datuak ditugu 3.3 taulan. Bertan, EGLUko informazioarekin osatu dugun egitura-zerrendatik EPEC-DEP corpusean zenbat egitura aurkitzen diren ikusi dugu.

Mota	Jokatuak	Ez-jokatuak
Denbora	40,00	60,00
Kausa	80,00	50,00
Kontzesioa	100,00	66,67
Modua	63,64	63,33
Helburua	100,00	33,33
Ondorioa	100,00	-
Baldintza	50,00	55,56

3.3 taula – Aurkitutako egituren agerpenak

Emaitzetan ikus daitekeenez, denbora-perpaus jokatuak eta helburu-perpaus ez-jokatuak dira corpusean barietate gutxien agertu duten egiturak. Kontzesio- eta ondorio-perpaus jokatuetan, berriz, EGLUn deskribatutako egitura guztiak agertu dira. Gehienak, ordea, % 60ren inguruan dabiltza. Gogoan izan behar da ondorio-perpaus jokatuetan egitura bakarra dagoela EGLUn eta, horrenbestez, % 100eko agerpena lortu du. Baldintza-perpaus jokatuetan, ordea, bi daude eta horietako bakarra aurkitu dugu. Datu horiekin lehendabiziko galdera erantzun dugu.

Bigarren galderari erantzunez, perpaus adberbial mota bakoitza corpusean zenbatetan agertzen den ikusi dugu. Maiztasun horiekin corpusaren banaketa nolakoa den jakin dugu. Perpaus adberbialen maiztasunak 3.4 tau-

lan ikus daitezke. Taularen bigarren zutabean corpusean perpaus adberbialen artean duten maiztasuna adierazi dugu. Hirugarren eta laugarren zutabeetan corpusean duten maiztasuna aditzaren arabera (jokatua edo ez-jokatua) eman dugu, eta bi horien baturak lehenengo zutabeko kopurua ematen du. Halaber, zutabe horietan parentesien artean agertzen diren kopuruekin, mota bakoitzean aditz jokatuen eta ez-jokatuen arteko banaketa adierazi dugu.

Perpaus-mota	**Guztira**	**Jokatuak**	**Ez-jokatuak**
Denbora	17,55	9,01 (51,34)	8,54 (48,66)
Kausa	17,10	16,61 (97,11)	0,49 (2,89)
Kontzesioa	5,76	3,01 (52,24)	2,75 (47,76)
Modua	29,95	5,86 (19,56)	24,09 (80,44)
Helburua	22,37	1,32 (5,89)	21,05 (94,11)
Ondorioa	0,28	0,28 (100,00)	-
Baldintza	6,99	5,86 (83,84)	1,13 (16,16)

3.4 taula – Perpaus-moten erabileraren maiztasunak

Emaitzetan ikus daitekeen bezala, nagusitzen diren perpausak modu- (% 29,95) eta helburu-perpausak (% 22,37) dira. Denbora- (% 17,55) eta kausa-perpausen (% 17,10) maiztasunak antzekoak dira, baita baldintza- (% 6,99) eta kontzesio- (% 5,76) perpausenak ere. Oso ondorio-perpaus gutxi dago (% 0,28).

Perpaus adberbialak aditzaren arabera (jokatua edo ez-jokatua) banatzen baditugu, motarik erabilienak dira modu ez-jokatuak (% 24,09) eta helburu ez-jokatuak (% 21,05); gutxien erabiliak dira helburu jokatuak (% 1,32), baldintza ez-jokatuak (% 1,13), kausa ez-jokatuak (% 0,49) eta ondorio jokatuak (% 0,28). Kausa-perpaus jokatuak (% 16,61) izan ezik, gainontzeko perpausak ez dira % 10era iristen.

Mota bakoitzaren aditzaren araberako sailkapena parentesien artean adierazi dugu. Mota batzuekin erabatekoa da aditz jokatua edo ez-jokatua erabiltzeko joera. Adibidez, kausa- (% 97,11) eta baldintza-perpausek (% 83,84) aditz jokaturako joera dute eta modu- (% 80,44) eta helburu-perpausek (% 94,11), aldiz, aditz ez-jokaturako. Ondorio-perpausak aditz jokatuarekin soilik erabiltzen dira. Denbora- eta kontzesio-perpausek, berriz, antzeko banaketa dute. Datu horiekin bigarren galdera erantzun dugu.

Hirugarren galderari erantzunez, perpaus adberbialen motek aditz nagusiarekiko duten kokapena 3.5 taulan azaldu dugu. Ikus daitekeen bezala, denbora-, kontzesio- eta modu-, eta baldintza-perpaus jokatuak aditz nagusiaren aurretik agertzen dira, eta kausa-, helburu- eta ondorio-perpausak,

berriz, ondotik.

Perpaus ez-jokatukatuei dagokienez, perpaus jokatuetan aurkitu ditugun joera beretsuak ikusi ditugu. Salbuespena modu-perpausak dira. Modu-perpausetan aditz jokatua dutenek aditzaren aurretik agertzeko joera daukate, baina ez-jokatua dutenek, berriz, aditzaren ondotik. Honenbestez, aditza jokatua izan ala ez, ez dirudi oro har aldaketa handirik dagoenik.

Mota	Aditza	Aurretik	Atzetik
Denbora	**Jokatuak**	67,27	32,73
	ez-jokatuak	74,19	25,81
Kausa	**Jokatuak**	29,49	70,51
	ez-jokatuak	37,88	62,12
Kontzesioa	**Jokatuak**	75,00	25,00
	ez-jokatuak	64,75	35,25
Modua	**Jokatuak**	63,29	36,71
	ez-jokatuak	33,63	66,37
Helburua	**Jokatuak**	14,29	85,71
	ez-jokatuak	41,74	58,26
Ondorioa	**Jokatuak**	14,29	85,71
Baldintza	**Jokatuak**	81,12	18,88
	ez-jokatuak	84,91	15,09

3.5 taula – Perpaus adberbial jokatuen eta ez-jokatuen kokapena aditz nagusiarekiko

Hirugarren galderaren erantzunarekin jakin dugu mendeko perpausek perpaus nagusiarekiko zein kokapen duten, alegia erlazio-informazioa ekintza nagusiarekiko non agertzen den. Datu horiekin, hurrenkera logikoak betetzen diren ala ez ikusi dugu. Kausa-perpausetan mendeko perpausa perpaus nagusiaren atzetik ageri denez, kausa-ondorio hurrenkera logikoa ez da betetzen. Dena den, sakonago aztertu behar dugu kausa azpimotaren arabera (beteak, azalpenezkoak edo inferientziazkoak) zein den jarraitzen den hurrenkera. Denbora-perpausetan ere sakonago aztertu behar dugu ea hurrenkera kronologikoa azpimotetan (aurrekotasuna, ondokotasuna...) betetzen den. Kontzesio-, modu-, helburu-, ondorio-, eta baldintza-perpausetan hurrenkera logikoak betetzen dira. Hurrenkera-datu horiek motaz mota ikusiko ditugu aurrerago. Informazio hori perpaus adberbialetatik sortuko ditugun esaldi berriak ordenatzeko erabiliko dugu.

Behin, corpusaren azterketa kuantitatiboarekin zein perpaus adberbial erabiltzen diren (agerpena), zenbatetan erabiltzen diren (maiztasuna) eta

nola erabiltzen diren (kokapena) ikusi ondoren, hurrengo azpiataletan perpaus adberbial bakoitzaren sinplifikazio-proposamenak azalduko ditugu datu horiek aintzat hartuta. Azalpenak ematerakoan, batez ere, txertatze-elementuetan, esaldi sinplifikatuen hurrenkeran eta ezaugarri berezietan erreparatu dugu.

Denbora-perpausak

Denbora-perpausek antolamendu kronologikoa adierazten dute. Hiru erlazio nagusi adierazten dituzte: aldiberekotasuna (orokorra, errepikatua eta estua), aldi desberdintasuna (aurrekotasuna eta ondokotasuna) eta iraupena (aurreko muga, azken muga eta tarte osoa). Mendeko denbora-perpausa da perpaus nagusiaren denbora edo kronologia zehazten duena.

Euskarazko egitura	Kopurua	Ehunekoa	Aurretik	Atzetik
-enean	223	58,22	65,92	34,08
-ela	72	18,80	83,33	16,67
-elarik	1	0,26	100,00	0,00
-enetan	1	0,26	100,00	0,00
-en bakoitzean	1	0,26	100,00	0,00
-eneko	2	0,52	100,00	0,00
-enerako	4	1,04	75,00	25,00
-en ostean	3	0,78	66,67	33,33
-enetik	51	13,32	54,90	45,10
-enez gero	2	0,52	50,00	50,00
-en arte	2	0,52	0,00	100,00
-en bitartean	21	5,48	61,90	38,10

3.6 taula – Denbora-perpaus jokatuak corpusean

Corpusean, denbora-perpausek perpaus adberbialen % 17,55 osatzen dute. Denbora-perpaus jokatu motaz motako emaitzak 3.6 taulan jaso ditugu, eta corpusean agerpenak dituzten egiturak soilik aurkeztu ditugu.

Denbora-perpaus jokatuen artean gehien erabili dena *-enean* da, % 58,22 maiztasunarekin. Ondoren, *-ela* eta *-enetik* egiturak dira erabilienak, % 18,80 eta % 13,32 maiztasunarekin, hurrenez hurren. % 5eko muga *-en bitartean* egiturak ere gainditu du. Lehendabiziko biek aldiberekotasuna adierazten dute, hirugarrenak iraupenaren aurreko muga eta azkenak iraupenaren tarte osoa.

Lau kasu horietan hurrenkera kronologikoa betetzen da, aditzaren aurretik ageri baitira gehienetan, *-enetik* egituraren banaketak ozta-ozta % 50eko

muga gainditzen duen arren. Ondokotasuna eta aurrekotasuna adierazten duten perpausen hurrenkeran ez dugu jakiterik izan hurrenkera kronologikoa betetzen den, 3 instantzia baino ez ditugulako aurkitu *-en ostean* erabiltzen dutenak eta ez dugulako aurkitu aurrekotasuna adierazten duen aditz jokatua duen egiturarik.

Euskarazko egitura	**Kopurua**	**Ehunekoa**	**Aurretik**	**Atzetik**
-tzean	59	13,29	77,97	22,03
-tzerakoan	24	5,38	66,67	33,33
-tzekoan	1	0,22	100,00	0,00
-tzearekin	1	0,22	0,00	100,00
-tu ahala	10	2,24	100,00	0,00
-tzerako	4	0,90	100,00	0,00
-tu orduko	9	2,02	88,89	11,11
-tu bezain laster	1	0,22	100,00	0,00
-tu bezain pronto	2	0,45	100,00	0,00
-tu eta berehala	1	0,22	100,00	0,00
-tu baino lehen	4	0,90	100,00	0,00
-tu aurretik	39	8,74	71,79	28,21
-tu eta	6	1,34	100,00	0,00
-tu eta gero	25	5,60	52,00	48,00
-tuta	4	0,90	75,00	25,00
-tutakoan	8	1,79	87,50	12,50
-tu ondoan	4	0,90	50,00	50,00
-tu ondoren	105	23,54	75,24	24,76
-tu ondotik	6	1,35	66,67	33,33
-tu ostean	83	18,61	80,72	19,28
-tuz gero	1	0,22	100,00	0,00
-tu arte	23	5,16	43,48	56,52
-tu artean	4	0,90	100,00	0,00
-tu bitartean	4	0,90	50,00	50,00
-tzeraino	1	0,22	0,00	100,00
-tu aitzin	4	1,90	75,00	25,00
-tu osteko	1	0,22	100,00	0,00
-tu bezperan	1	0,22	0,00	100,00
-tzear	1	0,22	100,00	0,00

3.7 taula – Denbora-perpaus ez-jokatuak corpusean

Denbora-perpaus ez-jokatuetan (3.7 taula) EPEC-DEP corpusean erabilienak *-tzean*, *-tzerakoan* (aldiberekotasuna); *-tu ahala* (aldiberekotasun errepikatua eta estua); *-tu orduko* (aldiberekotasun estua); *-tu aurretik* (aurrekotasuna); *-tu ondoren*, *-tu ostean* (ondokotasuna); eta *-tu arte* (iraupenaren

azken muga) dira. Horietatik egiturarik erabiliena *-tu ondoren* da % 23,54ko maiztasunarekin.

Denbora-perpaus ez-jokatuek esaldian duten kokapenari dagokionez, aldiberekotasuna (edozein) eta aurrekotasuna adierazten duten perpausek aditzaren aurretik agertzeko joera nabarmena dute. Iraupenaren azken muga adierazten duten perpausek aditzaren ostean agertzeko joera daukate. Ondokotasunak adierazten dutenak aditzaren aurrean joateko joera daukate *-tu eta gero*-ren kasuan instantzia gutxigatik den arren.

Aurrekotasuna eta ondokotasuna adierazten duten perpausen hurrenkerari erreparatuz, aurrekotasuna adierazten dutenek aditzaren aurretik azaltzeko joera handiagoa dute, ondokotasunekoek bezala. Baina aurrekotasuneko perpausen hurrenkera kronologikoaren aurka doa, hau da, lehenengo ageri den ekintza, denbora-lerroan ondoren gertatzen da.

Behin corpusean aztertuta zein diren denbora-perpausik usuenak eta horien kokapena, sinplifikazio-proposamenak deskribatzeari ekingo diogu. Perpaus-mota bakoitzak bere berezitasunak dituen arren, **denbora perpausen sinplifikazio-proposamenak** sistematizatzeko helburuarekin esanahiaren arabera sailkatu ditugu (EGLUko sailkapenari jarraituz).

Aldiberekotasuna (17) adierazten dutenekin hasiko gara. Honako hau da sinplifikazio-proposamena:

1. Esaldian dauden perpausak banatu: [Edurnezuri printzearekin ezkondu zenean,]$_{p_1}$ [zazpi ipotxek edateari eman zioten.]$_{p_2}$

2. Mendeko perpausetik erlazio-marka ezabatu: zenean -> zen

3. Perpaus nagusian aldiberekotasuna adierazten duen adberbio bat (*Orduan*[13]) txertatu. TE alternatiboak *Une hartan* edo *Aldi berean*[14] dira.

4. mendekoa$_{jat}$-nagusia$_{jat}$ ordenatu

5. Puntuazio-markak egokitu

(17) a. Edurnezuri printzearekin ezkondu ze**nean**, zazpi ipotxek edateari eman zioten.

[13]Lemen maiztasun-zerrendan 19.892 agerpen ditu "*Orduan*"-ek eta, beraz, egokia dela uste dugu *-enean* egiturak duen inesibo kasu-marka mantentzen duelako. Ahal izan den kasu guztietan *ordu + egituraren kasu-marka* duen TEak proposatu ditugu.

[14]Maiztasunak: *une* 14.710; *aldi berean* 2.595.

b. i. Edurnezuri printzearekin ezkondu zen.
ii. *Orduan* zazpi ipotxek edateari eman zioten.

Aditz ez-jokatuak dituzten perpausekin (18) egiten dugun sinplifikazio-proposamena antzekoa da. Ezberdintasuna da aditz ez-jokatua jokatu bihurtu behar dugula.

(18) a. Bill Clinton presidentearekin bat etorri zen atzo Janet Reno fiskal nagusia Elian Gonzalezen kasuari buruz hitz egi**terakoan**.
b. i. Elian Gonzalezen kasuari buruz hitz egin dute.
ii. *Orduan* Bill Clinton presidentearekin bat etorri zen atzo Janet Reno fiskal nagusia.

Aldiberekotasun orokorra adierazten duten perpausak izan ditugu aztergai (17) eta (18) adibideetan, baina aldiberekotasun errepikatua eta estua adierazten duten perpausen sinplifikazio-proposamena antzekoa da. Txertatze-elementuak dira soilik aldatuko direnak. Aldiberekotasun errepikatuan hautatutako TEa *Une horietan guztietan* da eta alternatiboa *Aldiro*. Kasu honetan sintagma txertatzeko aukera hobetsi dugu sintagmaren buruak *Une Aldiro* elementuak baino maiztasun[15] altuagoa duelako. Aldiberekotasun estuan, *Orduko*[16] eta *Une horretan bertan* aukeratu ditugu eta alternatiboa *Segidan*[17] da. TE bakoitza zein egiturari dagokion 3.8 taulan jaso dugu, denbora-egituretan TEak laburbiltzen dituen taulan, alegia.

Aldi ezberdintasuna (aurrekotasuna eta ondokotasuna) adierazten duten egiturei erreparatuko diegu orain. Aurrekotasuna adierazten duten perpausetan, aukeratu dugun TEa *Gero* da eta alternatiboak *Ondoren* eta *Ostean*[18]. Ondokotasuna adierazten dutenetan, berriz, *Ondoren* TEa eta *Ostean* alternatiboa. Aditz ez-jokatua duten perpausetan, aditz jokatuak sortu ditugu.

Hurrenkerari helduz, corpusean hurrenkera kronologikoa mantentzen ez dela ikusi dugun arren, gure ustez eta sinplifikazioko beste lanetan ikusten den lez (Specia *et al.*, 2008; Klerke eta Søgaard, 2012), hurrenkera kronologikoa errespetatzeak errazagoa egiten du ulermena. Erabaki hori zergatik hartu dugun (19) adibidearekin azalduko dugu. Adibideetako b) multzoko

[15] Maiztasunak: *une* 14.710; *aldiro* 280.
[16] *-eneko* = menderagailua + leku genitiboa; beraz, txertatze-elementua *ordu* + leku genitiboa.
[17] Maiztasunak: *une* 14.710; *orduko* 14.268; *segidan* 535.
[18] Maiztasunak: *gero* 69.726; *ondoren* 42.689; *ostean* 7.452.

esaldiak ikusita, agindu bat adierazten da (19b-i), baina (19b-ii) esaldia irakurrita lehenago gertatu den edo behar den gauza bat adierazten da. Kasu horietan gaizki ulertuak egon daitezkeela iruditu zaigu. Adibideetako c) multzoko esaldiak irakurrita, aldiz, zein hurrenkera kronologikotan gertatu diren edo gertatu behar diren ondo ulertzen dela iruditu zaigu.

(19) a. Jarduera honekin has**i aurretik** egizu aurrekoan irakurritako testuaren laburpen bat.

b. i. Jarduera honekin hasi.

ii. *Aurretik* egizu aurrekoan irakurritako testuaren laburpen bat.

c. i. Egizu aurrekoan irakurritako testuaren laburpen bat

ii. *Gero*, jarduera honekin hasi.

Ondokotasunaren kasuan (20), esaldien hurrenkera corpuseko datuek adierazi bezala eman dugu, kasu horretan hurrenkera kronologikoa mantentzen baita.

(20) a. Sei hilabetetan isilik egonø **ostean**, Marcosek historikotzat jo zuen PRIren aroaren amaiera.

b. i. Sei hilabetetan isilik egon zen.

ii. *Ondoren*, Marcosek historikotzat jo zuen PRIren aroaren amaiera.

Iraupena adierazten duten perpausei (21) helduko diegu orain. Iraupenaren aurreko muga adierazten duten perpausetan *Ordutik*[19] da TEa eta *Harrezkero*[20] da alternatiboa. Iraupenaren azken muga adierazten duten perpausetan TEa *Ordura arte* da eta *Orduraino*[21] alternatiboa. Iraupenaren tarte osoa adierazten dutenenetan, berriz, *Bitartean* da TEa eta alternatiboa *Artean*[22], maiztasun-zerrendan begiratuta *artean* hitzak maiztasun altuagoa duen arren[23]. Hurrenkera dela-eta, corpuseko datuekin ez dugu patroi garbirik lortu, batez ere instantzia gutxi daudelako. Instantzia-kopuru altuagoak

[19] *-enetik* = menderagailua + ablatiboa; beraz, TEa *ordu* + ablatiboa.

[20] Maiztasunak: *ordutik* 1.737; *harrezkero* 522.

[21] Maiztasunak: *ordura arte* 996; *orduraino* 7.

[22] Maiztasunak: *bitartean* 15.712; *artean* 20.482.

[23] Maiztasun-zerrendan lemak soilik hartu dira kontuan eta ez kategoriak; beraz, ezin dugu bereizi zenbat agerpen dagozkien postposizioei eta zenbat adberbioei. Guk *Bitartean* aukeratu dugu *-en eta -tu bitartean* egituren baturaren emaitza handiagoa delako *-en eta -tu artean* egiturena baino.

dituzten egituretan ere ezin da joera bat ezarri. Beraz, denbora-perpausek oro har aurretik joateko joera dutenez, mendekoa nagusiaren aurretik ematea erabaki dugu.

(21) a. Zuen erantzuna jaso genue**netik** jokabide guztiz ezberdinak antzeman ditugu zuen alderdikideen artean.

b. i. Zuen erantzuna jaso genuen.

ii. *Ordutik*, jokabide guztiz ezberdinak antzeman ditugu zuen alderdikideen artean.

Esaldietan gehitu ditugun TEak, TE alternatiboak eta dagozkien egiturak 3.8 taulan laburbildu ditugu. Esaldi berrien hurrenkera ere taula berean jaso dugu. Izartxoa duten egiturak multzo batean baino gehiagotan sailkatu ditugu.

Multzoa	**Egitura**	**Txertatze-elementua**	**Txertatze-elementu alternatiboak**	**Hurrenkera**
Aldiberekotasun orokorra	*-enean; -ela(rik); noiz eta ... bait- /-en*	Orduan$_{nag}$	Une hartan$_{nag}$; Aldi berean$_{nag}$	mendekoa$_{jat}$-nagusia$_{jat}$
	-tzean; -tzerakoan; -tzekoan; -tzearekin; -tzeari/ -tzerat; -tu(k)eran			
Aldiberekotasun errepikatua	*-enetan; -en bakoitzean; -en guztietan; -en aldikal/aldiro; zenbat aldiz -en ... hainbat aldiz, -tu aldiro; -tu bakoitzean; -tu guztian; -tu ahala/arau*	Une horietan guztietan$_{nag}$	Aldiro$_{nag}$	mendekoa$_{jat}$-nagusia$_{jat}$
Aldiberekotasun estua	*-eneko; -en orduko*	Orduko$_{nag}$	Segidan$_{nag}$	mendekoa$_{jat}$-nagusia$_{jat}$
	-tzerako; -tu orduko			
	-en bezain laster/sarri/agudo/fite; -en ber; -enaz batera	Une horretan bertan$_{nag}$	Orduko$_{nag}$; Segidan$_{nag}$	

(Jarraipena hurrengo orrialdean)

Multzoa	**Egitura**	**Txertatze-elementua**	**Txertatze-elementu alternatiboak**	**Hurrenkera**
	-tu bezain laster/pronto; -tu eta berehala; -tu eta laster; -tuaz/-tzearekin bat(era); -tu berri(t)an; -tu ahala/arau			
Aurrekotasuna	*-en baino lehen; -en aurrean/aitzinean* *-tu baino lehen; -tu aurretik/ -tu aitzinean; -tu gabe; -tu orduko*; -tzerako**	Gero$_{nag}$	Ondoren$_{nag}$; Ostean$_{nag}$	nagusia$_{jat}$-mendekoa$_{jat}$
Ondokotasuna	*-en ondoan; -en ondoren; -en ostean* *-tu eta; -tu eta gero; -tuta; -tu ondoan; -tu ostean; -tu(a)z; -tuz gero; -turik*	Ondoren$_{nag}$	Ostean$_{nag}$	mendekoa$_{jat}$-nagusia$_{jat}$
Iraupenaren aurreko muga	*-enetik; -enez gero; -enik ...-ra* *-tuz gero**	Ordutik$_{nag}$	Une hartatik$_{nag}$; Harrezkero$_{nag}$	mendekoa$_{jat}$-nagusia$_{jat}$
Iraupenaren azken muga	*-en arte* *-tu arte; -tu artean; -tu bitartean; -tzeraino*	Ordura arte$_{nag}$	Orduraino$_{nag}$	mendekoa$_{jat}$-nagusia$_{jat}$
Iraupenaren tarte osoa	*-eno/-eino; -en bitartean; -en artean; -en arteko* *-tu bitarte(an)*; -tu artean**	Bitartean$_{nag}$	Artean$_{nag}$	mendekoa$_{jat}$-nagusia$_{jat}$

3.8 taula – Denbora-perpausen txertatze-elementuak eta sortutako esaldi berrien hurrenkera

Bestalde, EPEC-DEP corpuseko azterketan aurkitu ez diren egiturak edo gutxitan agertu direnak sinplifikatzeko beste sinplifikazio-mota bat proposatu dugu: ordezkapen sintaktikoen sinplifikazioa. Sinplifikazio-mota horren adibide bat (22) da. Adibide horretan ageri den *-tzeari* egitura corpusean aurkitu ez dugunez, aldiberekotasun orokorra ere adierazten duen *-tzean* egiturarekin ordezkatu dugu. Adibidea EGLUtik atera dugu, gure corpusetan egitura hori ez baitugu aurkitu.

(22) a. Zahar**tzeari**, boza galduz geroz, ezin aditua zen, eta (...).
b. i. Zahar*tzean*, boza galduz geroz, ezin aditua zen, eta (...).

Aditz ez-jokatua duten denbora-perpausekin proposatzen ditugun ordezkapenak 3.9 taulan jaso ditugu.

Maiztasun gutxiko egiturak	**Ordezkapenerako hautagaia**
-tzeari; -tzera(t); -tu(k)eran	-tzean
-tu arau	-tu ahala
-tu baino lehen; -tu gabe; -tu aintzinean	-tu aurretik
-tu eta; -tu eta gero; -tu ondoan; -tuz gero; -tuaz gero; -tu ostean; -tu ondotik	-tu ondoren
-tu artean	-tu arte

3.9 taula – Maiztasun gutxiko denbora-perpausen egitura ez-jokatuak ordezkatzeko proposamenak

Kausa-perpausak

Kausa-perpausek perpaus nagusiko ekintzaren kausa edo zergatia azaltzen dute. Kausa-perpausen artean kausal beteak, azalpenezkoak eta inferentziazko azalpenezkoak sailkatu ohi dira; lokailuekin sortzen direnak eta bestelakoak ere badaude. Guk hasierako hiru multzotakoak landuko ditugu, baina azalpenezkoak eta inferentziazko azalpenezkoak batera emango ditugu, zaila baita bien artean bereiztea. Azken finean, biak azalpenak dira eta egiturak errepikatu egiten dira. EPEC-DEP corpuseko kausa-egituren azterketan ikusi dugu perpaus adberbialen % 17,10a osatzen dutela. 3.10 taulan corpusean agertu diren egitura bakoitzaren maiztasunak eta aditz nagusiarekiko kokapena aurkezten dugu.

Euskarazko egitura	**Kopurua**	**Ehunekoa**	**Aurretik**	**Atzetik**
-elako((t)z)	190	26,91	11,05	88,95
bait-	282	39,94	6,38	93,62
zeren eta...(bait-/-(e)n)	4	0,57	0,00	100,00
... eta	51	7,22	74,51	25,49
-tzeagatik	2	9,52	50,00	50,00
-tzearren	2	9,52	50,00	50,00

3.10 taula – Kausa-perpausak corpusean

Aditz jokatua duen egiturarik erabiliena *bait-* izan da (% 39,94); egitu-

ra hori kausal betea, azalpenezkoa eta inferentziazkoa da. Ez-jokatuetako bi egiturek (*-tzeagatik* eta *-tzearren*) maiztasun bera izan dute (% 9,52) eta gutxitan agertu dira (bi aldiz). Kokapenari dagokionez, ... *eta*[24] egitura izan ezik, gainontzeko guztiek perpaus nagusiaren atzetik agertzeko joera dute. Beraz, hurrenkera logikoa (kausa-ondorioa) betetzen ez dela ondoriozta dezakegu. Aditz ez-jokatua duten perpausen kasuan ezin da ezer ondorioztatu; perpaus gutxi aurkitzeaz gain, bakoitzak bere posizioa du.

Kausa-perpausen sinplifikazio-proposamenei dagokienez, kausal beteak (23) sintaktikoki sinplifikatzeko perpaus nagusian TEa *Horregatik*[25] da eta TE alternatiboa *Hori dela-eta*[26]. Corpusean kausal beteak perpaus nagusiaren ondotik agertzeko joera argia duten arren, guk kausa-ondorio hurrenkera logikoa jarraitzea proposatu dugu, sinplifikazioko beste lan batzuetan agertzen den bezala (Specia *et al.*, 2008). Hori dela-eta, esaldi berriak mendekoa$_{jat}$-nagusia$_{jat}$ hurrenkerari jarraituz eman ditugu.

(23) a. Lizarraren sinatzaileekiko autokritika ere egin du ELAko idazkariak, "akordioa ekintza politiko eta sozialean gauzatzeko" gauza izan ez dire**lako**.

b. i. Ez dira "akordioa ekintza politiko eta sozialean gauzatzeko" gauza izan.

ii. *Horregatik* Lizarraren sinatzaileekiko autokritika ere egin du ELAko idazkariak.

Aditz ez-jokatua duten kausa-perpausekin ere kasual beteekin bezala egin dugu. Gogoan izan behar dugu aditz ez-jokatua duten denbora-perpausen kasuan bezala, aditz jokatuak sortu behar ditugula.

Azalpenezko kausa-perpausen kasuan (24), ordea, TEa *Izan ere*[27] da eta mendeko perpausean txertatu dugu. Erabaki hori hartu dugu kasu honetan kausa-ondorio elkarrekintza agertzen ez delako, baizik eta azalpen bat; hori dela-eta, erabaki dugu azalpenak gertatu den ekintzaren ondotik eman behar direla. Gainera, hori izan da corpusean aurkitu dugun joera.

[24]Egitura hori sarrera-egitura bezala erabiltzen da maiz.

[25]Maiztasuna: *horregatik* 9.470.

[26]Antolatzaile honen egitura dela-eta, ezin izan dugu bere maiztasuna maiztasun-zerrendan konprobatu.

[27]Maiztasuna: *izan ere* 13.686.

(24) a. Ez zen ustekabeko handirik izan atzo Kontxarako sailkatzeko estropadan, ezin **bait**a ezustekotzat jo Orio, Trintxerpe, Tiran, Donibaneko, Astillero, Koxtape eta Hondarribia sailkatu izana.

b. i. Ez zen ustekabeko handirik izan atzo Kontxarako sailkatzeko estropadan.

ii. *Izan ere*, ezin da ezustekotzat jo Orio, Trintxerpe, Tiran, Donibaneko, Astillero, Koxtape eta Hondarribia sailkatu izana.

Kausa-perpausetan gehituko ditugun TEak eta esaldi berrien hurrenkera 3.11 taulan jaso ditugu.

	Egitura	**Txertatze-elementua**	**Txertatze-elementu alternatiboak**	**Hurrenkera**
Kausal beteak	*-elako/ -elakoz/ -lakotz; -elakoan; bait-; zeren eta ... bait-/ -(e)n -tzeagatik; -tzearren*	Horregatik$_{nag}$	Hori dela-eta$_{nag}$	mendekoa$_{jat}$-nagusia$_{jat}$
Azalpenezkoak	*Bait-; ... eta; zeren eta ... (bait-/-(e)n)*	Izan ere$_{men}$	-	nagusia$_{jat}$-mendekoa$_{jat}$

3.11 taula – Kausa-perpausen txertatze-elementuak eta sortutako esaldi berrien hurrenkera

Ordezkapen sintaktikoen sinplifikazioa egiteko, kausa-perpausen kasuan bi egiturek maiztasun bera dute, baina *-tzearren* egitura anbiguoa denez (helburua ere adieraz dezake) *-tzeagatik* aukeratu dugu ordezkapenak egiteko.

Kontzesio-perpausak

Kontzesio-perpausek semantikoki oztopo bat edo bateraezintasuna adierazten dute perpaus nagusian azaltzen den ekintzarekiko. Corpuseko perpaus adberbialen % 5,76 dira. 3.12 taulan adierazi ditugu kontzesio-perpausen egituren maiztasunak eta kokapenak.

Aditz jokatua dutenen artean, egiturarik erabilienak *ba- (...) ere* (% 44,53) eta *-en arren* (% 43,75) dira, eta biek aditz nagusiaren aurretik agertzeko joera dute. Kontzesio-perpaus ez-jokatuetan gehien erabiltzen diren egiturak *-tu*

Euskarazko egitura	Kopurua	Ehunekoa	Aurretik	Atzetik
nahiz (eta) (-en/-ela/ba-)	15	11,72	40,00	60,00
-en arren	56	43,75	76,79	23,21
ba- (...) ere	57	44,53	85,96	14,04
nahiz eta... /-tu/ø	28	23,93	39,29	60,71
-tu arren	77	65,81	76,62	23,38
-tuagatik	1	0,85	100,00	0,00
-tuta (gabe/ezta) ere	2	1,71	50,00	50,00
-tzearren	5	4,27	40,00	60,00
-tuz gero	2	1,71	100,00	0,00
-ik ere	2	1,71	100,00	0,00

3.12 taula – Kontzesio-perpausak corpusean

arren (% 65,81) eta *nahiz eta... /-tu/ø* (% 23,93) dira, eta horiek ere aurretik agertzeko joera dute.

Kontzesio-perpausak (25) **sinplifikatzean** ez dugu azpisailkapenik egin eta sinplifikazio sintaktikoa egitean egitura guztietan *Hala ere* txertatu dugu perpaus nagusian. TE alternatiboak *Nolanahi ere, Edonola ere* eta *Hala eta guztiz ere*[28] dira. Hurrenkera mendekoa$_{jat}$-nagusia$_{jat}$ da, corpuseko datuetan agertzen den bezala.

(25) a. Hasiera batean aste honetan partidurik ez jokatzea aurreikusita zego**en arren**, azken orduan ostiralean Holandara abiatu aurretik partidu bat jokatu nahi izan du Alavesek.

b. i. Hasiera batean aste honetan partidurik ez jokatzea aurreikusita zegoen.

ii. *Hala ere*, azken orduan ostiralean Holandara abiatu aurretik partidu bat jokatu nahi izan du Alavesek.

Ordezkapen sintaktikoen sinplifikazioan *-tu arren* egiturarekin ordezkatu ditugu maiztasun gutxiko egiturak.

Modu-perpausak

Modu-perpausek aditz nagusiko ekintza nola betetzen den adierazten dute. Corpusean perpaus adberbialen % 29,95 osatzen dute eta perpaus adberbialetako motarik erabiliena da. Modu-perpaus jokatuetan (3.13 taula) gehien erabiltzen den egitura *-enez* da, % 49,80ko maiztasuna duena. Erabilienen

[28]Maiztasunak: *hala ere* 11.615; *nolanahi ere* 1.505; *edonola* 933; *hala eta guztiz ere* 595.

artean *-elarik* (% 18,88), *-ela* (% 18,07) eta *-en bezala* (% 11,65) ditugu. *-elarik* egituraren kausan izan ezik guztiek aditzaren aurretik azaltzeko joera daukate.

Euskarazko egitura	Kopurua	Ehunekoa	Aurretik	Atzetik
-ela	45	18,07	73,33	26,67
-elarik	47	18,88	23,40	76,60
-en moduan/ra	1	0,40	100,00	0,00
-en antzera	1	0,40	0,00	100,00
-en bezala	29	11,65	68,97	31,03
-en legez	2	0,80	0,00	100,00
-enez	124	49,80	76,61	23,39

3.13 taula – Modu-perpaus jokatuak corpusean

Modu-perpaus ez-jokatuek (3.14 taula) egitura anitz dituzte, eta, EGLUtik sortutako zerrendan daudenez gain, taulan letra etzanez jarri ditugun beste sei egitura aurkitu ditugu corpusean, maiztasun handikoak, gainera. Egitura horiek ziurrenik corpusa etiketatzean sortutako erroreak dira, eta azterketa honetan, corpuseko errore horiek antzemateaz gain egitura horiek nahasgarriak direla eta kontuan hartu beharrekoa direla iruditu zaizkigu.

Emaitzak analizatuz, aditz ez-jokatua duten egiturarik erabilienak *-tuta* eta *-tuz* dira, % 41,61eko eta % 33,31ko maiztasunarekin, hurrenez hurren. Gainontzeko egiturak ez dira % 10eko mugara iristen eta horietatik 10 ezta % 1era ere. Perpaus horien arteko hurrenkerari dagokionez, aditzaren aurrean agertzeko joera duten egiturarik erabilienak *-tuta*, *-turik*, *-tu ordez*, *-tu ahala*, *-tu beharrean*, *-tzeko moduan*, *-tzera*, *-tzeaz*, *-kotan*, *-tzeaz gain*, *-tik* eta *-tu bezala* dira eta aditzaren ostean, berriz, *-tuz*, *-tu gabe* eta *-tu nahian*. Beraz, hurrenkera orokortzea zaila egin zaigu.

Euskarazko egitura	Kopurua	Ehunekoa	Aurretik	Atzetik
-tuz	401	33,31	48,13	51,87
-tuta	501	41,61	61,68	38,32
-turik	86	7,14	63,95	36,05
-tu gabe	59	4,90	38,98	61,02
-tu barik	2	0,17	50,00	50,00
-tu ordez	5	0,42	80,00	20,00
-tu ahala	5	0,42	80,00	20,00
-tu beharrean	7	0,58	85,71	14,29
-tu nahirik	1	0,08	0,00	100,00
-tu nahian	8	0,66	25,00	75,00

(Jarraipena hurrengo orrialdean)

Euskarazko egitura	Kopurua	Ehunekoa	Aurretik	Atzetik
-tu ezinik	1	0,08	0,00	100,00
-tu ezinda	3	0,25	33,33	66,67
-tu ezinean	1	0,08	100,00	0,00
-tzeko zorian	3	0,25	33,33	66,67
-tzeko moduan	8	0,66	87,50	12,50
-tzeko eran	1	0,08	100,00	0,00
-tzeke	2	0,17	50,00	50,00
-tzekotan	8	0,66	87,50	12,50
-tu bezala	20	1,66	65,00	35,00
-tzera	41	3,41	92,68	7,32
-tzeaz	16	1,33	87,50	12,50
-tzeaz gain	19	1,58	100,00	0,00
-tik	6	0,50	100,00	0,00

3.14 taula – Modu-perpaus ez-jokatuak corpusean

Modu-perpausen sinplifikazio-proposamenean, jokatuetan *Hala* txertatu dugu eta TE alternatiboak *Honela, Horrela, Modu horretan, Era berean* eta *Era horretan* dira[29]. Esaldi berriak mendekoa$_{jat}$-nagusia$_{jat}$ hurrenkeran ordenatu ditugu. *-enez* egituraren kasuan aditz diskurtsiboak ez dituztenak sinplifikatuko ditugu modu horretan.

(26) a. ezkongaiaren gurasoek dirutan edo ondasunetan etxera ezkontzen den beste ezkongaiari emandako kopurua da, ezkongai horrek jasotzen dituen etxearen zamei erantzuteko balio du**elarik**.

b. i. Ezkongai horrek jasotzen dituen etxearen zamei erantzuteko balio du.

ii. *Hala*, ezkongaiaren gurasoek dirutan edo ondasunetan etxera ezkontzen den beste ezkongaiari emandako kopurua da.

Aditz ez-jokatua duten modu-perpausei dagokienez (27), jokatuetan bezala, esanahiak kontrakoa eskatu ezean, *Hala* da TEa eta alternatiboak *Honela/Horrela/Era berean/Modu horretan/Era horretan.* Aditz jokatua duten perpausekin koherentzia mantentzeko, mendekoa$_{jat}$-nagusia$_{jat}$ hurrenkera erabili dugu, corpusean alderantzizko hurrenkera gertatzen dela ikusi dugun arren.

(27) a. Demokritok, bi filosofo hauei aurre egin**ez**, beste hipotesi hau bota zuen:

[29]Maiztasunak: *hala* 24.220; *horrela* 8.178; *modu* 28.514; *era* 14.725.

b. i. Demokritok bi filosofoei aurre egin zien.
ii. *Hala*, Demokritok beste hipotesi hau bota zuen:

Egitura jakin batzuetan sinplifikazio-proposamenak egiteko, txertatzeko ohiko elementuez gain (lehenetsiak eta alternatiboak), txertatzeko bestelako elementuak erabili ditugu. Elementu horiek egituren esanahiarekin eta formarekin lotuta daude, eta horiek aukeratzeko ez dugu maiztasun-zerrenda erabili. Txertatzeko bestelako elementuak zein diren azalduko ditugu hurrengo lerroetan.

-tu nahirik, *-tu ezinik* eta *-tu beharrean/beharrez egon* egiturak sinplifikatzeko (28), aditz modalak gehitu ditugu mendeko perpausetan. Horrela, egitura horien ordez, *nahi izan*, *ezin izan* eta *behar izan* aditz modalak txertatu ditugu hurrenez hurren, nagusiaren aldi eta pertsona berean jokatuz.

(28) a. Eta gazte horien ilusioaren adinakoa da Al Goreren aholkularien kezka, bi aldeak (zentro eskuina eta aurrerazaleak) aseko dituen mezua topa**tu ezinik**.
b. i. Eta gazte horien ilusioaren adinakoa da Al Goreren aholkularien kezka.
ii. *Hala*, *ezin dute* bi aldeak (zentro eskuina eta aurrerazaleak) aseko dituen mezua topatu.

Ezezko kutsua duten modu-perpaus ez-jokatuak (29) sinplifikatzean, mendeko perpausa ezezkoan jarri dugu (barne-hurrenkera ere aldatuz) jatorrizko esaldiaren esanahiari eusteko. Horrelako egiturak dira *-tu gabe/barik/ezta*, *-tu ordez/ordean*, *-tu beharrean* eta *-tzeke*.

(29) a. Belgiarrei aitorturiko eskubide-askatasunak erabiltzea bermatzen da, bereizkeriarik egin**ø gabe**.
b. i. *Ez* da bereizkeriarik egiten.
ii. *Hala*, Belgiarrei aitorturiko eskubide-askatasunak erabiltzea bermatzen da.

-tu aginean/aginik, *-tu hurran* eta *-tzeko zorian* egiturek gertatzear dauden gertaerak azaltzen dituzte. Horiek sinplifikatzean, mendeko perpausean *ia* txertatzeko proposamena egin dugu. Corpusean honelako adibiderik topatu ez dugunez, EGLUko adibidea (30) baliatu dugu.

(30) a. (...) luze ibili ginen gal**du aginean**.

b. i. *Ia* galdu ginen.
ii. *Hala* luze ibili ginen.

Ekintza baten prozesua adierazten duten egituretan (*-tu ahala/arau*, (31) adibidea[30]) *-ten* aspektu-marka gehitu diogu mendeko perpausean sortzen dugun aditzari. Horrekin prozesua adierazi nahi izan dugu.

(31) a. Bizkarraldean zeuken pilloa agor**tu-ala**, (...) eramaten zitun neskak (...).
b. i. Bizkarraldean zeuken piloa agortzen zen.
ii. *Hala*, (...) eramaten zituen neskak (...).

-tzeko moduan, -tzeko gisan, -tzeko eran, -tzeko maneran, -tzeko moldean eta *-tu bezala* egiturak ez ditugu sinplifikatuko, *modu, gisa, era, manera, molde* eta *bezala* hitzetan jada modua adierazten delako. *-tzekotan* ere ez dugu modu-perpausak bezala sinplifikatu.

Goiko analisietan proposatu ditugun txertatzeko bestelako elementuak 3.15 taulan laburbildu ditugu.

Egitura	**Txertatzeko bestelako elementuak**
-tu nahirik; -tu ezinik; -tu beharrean/beharrez	nahi izan; ezin izan; behar izan
-tu gabe/barik/ezta;-tzeke; -tu ordez/ordean; -tu beharrean	ez (ezezkoan jarri)
-tu aginean/aginik, -tu hurran; -tzeko zorian	ia
-tu ahala/arau	*-ten* aspektua

3.15 taula – Modu-perpaus ez-jokatuetan txertatzeko bestelako elementuak

Ordezkapen sintaktikoen sinplifikazioko baliokidetzak 3.16 taulan ikus daitezke eta (32) da sinplifikazio-mota honen adibide.

Maiztasun gutxiko egiturak	**Ordezkapenerako hautagaia**
-tu arau	-tu ahala
-tu gurarik	-tu nahian

(Jarraipena hurrengo orrialdean)

[30] Adibide hori EGLUtik (Euskaltzaindia, 2011) atera dugu, EPECen daudenak luzera minimoa betetzen ez dutelako. EGLUn erreferentzia hau ematen da: (Agirre, Uztaro, 159).

Maiztasun gutxiko egiturak	**Ordezkapenerako hautagaia**
-tu barik; -tu ezta	-tu gabe
-tu ordean	-tu ordez
-tzeko gisan; -tzeko maneran	-tzeko moduan
-tu hurran; -tu aginean; -tu aginik;	-tzeko zorian

3.16 taula – Maiztasun gutxiko modu-perpausen egitura ez-jokatuak ordezkatzeko proposamenak

(32) a. Zabortegia egin**ø beharrean**, turismoa erakar dezaketen proiektuak bultzatu beharko genituzke, beharbada, Nobel Sarien arrakasta aprobetxatuz.

b. i. Zabortegia egin*ø ordez*, turismoa erakar dezaketen proiektuak bultzatu beharko genituzke, beharbada, Nobel Sarien arrakasta aprobetxatuz.

Orain arteko perpaus adberbialekin erkatuz, modu-perpausen sinplifikazio-proposamenen berezitasuna da erlazioa mantentzeko, txertatzeko bestelako elementuak proposatu ditugula aditz ez-jokatua duten zenbait egituratan.

Ondorio-perpausak

Ondorio-perpausek edo perpaus kontsekutiboek perpaus nagusian gertatzen den ekintzaren ondorioa adierazten dute. Ondorio hori ez da borondatezkoa eta kuantifikatzaile batekin erlazionatuta eman ohi da. Corpusean perpaus adberbialen % 0,28 osatzen dute. Oso instantzia gutxi dauden arren, corpusean egiturarik erabiliena *(...) (non) ... bait-* da, % 58,33ko maiztasunarekin eta aurkitutako bi egiturek aditz nagusiaren ondotik joateko joera dute (3.17 taula).

Euskarazko egitura	**Kopurua**	**Ehunekoa**	**Aurretik**	**Atzetik**
mailakatzailea (non/ezen) ... bait-	7	58,33	14,29	85,71
mailakatzailea (non/ezen) ... -en	5	41,67	0,00	100,00

3.17 taula – Ondorio-perpaus jokatuak corpusean

Ondorio-perpausak sinplifikatzeko (33), erlazio-marka ezabatzeaz gain, *non* eta *ezen* lokailuak ezabatu ditugu. TEa *ondorioz* da eta alternatiboak *beraz*, *hortaz* eta *honenbestez*[31] dira. Ondorio-perpausetan perpaus

[31]Maiztasunak: *ondorio* 17.734; *beraz* 10.698; *hortaz* 2.859; *honenbestez* 106.

Jatorrizko perpauseko kuantifikatzailea	Sinplifikatutako ordezkoa
hain	oso
hainbeste	asko
hainbat	hainbat
hala	hala, honela
halako maneran	hala, honela
halako modez	hala, honela
halako x	halako x

3.18 taula – Kuantifikatzaileak jatorrizko perpausean eta dagokien baliokidea sinplifikatutatako perpausean

nagusiko kuantifikatzailea ere ordezkatu dugu. Esaldi berrien hurrenkera nagusia$_{jat}$-mendekoa$_{jat}$ da, hori baita hurrenkera logiko eta kronologikoa (kausa-ondorioa) eta corpus-azterketan aurkitutakoa. Baliokidetza horiek 3.18 taulan aurkezten ditugu.

(33) a. Hain asaldaturik zegoen, **non** gelan gora eta behera **bait**zebilen, gauzak batetik bestera aldatzen, arrazoirik gabe.

b. i. *Oso* asaldaturik zegoen.

ii. *Ondorioz*, gelan gora eta behera zebilen, gauzak batetik bestera aldatzen, arrazoirik gabe.

Beste perpaus adberbialekin erkatuz, ondorio-perpausen berezitasuna kuantifikatzaileen aldaketa da.

Helburu-perpausak

Helburu-perpausek *zertarako* galdera erantzuten dute eta helburu bat adierazten dute, edo beste hitz batzuekin esanda, borondatezko ondorio bat. Ondorio-perpausekin duten ezberdintasunik handiena semantika aldetik horixe da hain zuzen, ondorioa borondatezkoa dela eta ondorio-perpausetan, aldiz, ez-borondatezkoa. Corpuseko % 22,37 osatzen dute eta horietatik % 94,11 ez-jokatuak dira.

Helburu-perpaus ez-jokaturik erabiliena *-tzeko(tz)* da (% 88,38). EPEC-DEPen gaizki etiketatuta dagoen *-tzera* egituraz gain, (% 9,83) gainontzeko egiturak ez dira % 2ra iristen. Helburu-perpausek aditzaren ondoren joateko joera dute. Joera hori perpaus jokatuekin ez ezik, ez-jokatuetako erabilienetako birekin ere (*-tzeko* eta *-tzearren*) mantentzen da. Emaitzak 3.19 taulan ikus daitezke.

Euskarazko egitura	Kopurua	Ehunekoa	Aurretik	Atzetik
-n (subjuntiboa)	56	100,00	17,86	82,14
-tzeko(tz)	791	88,38	40,20	59,80
-tzearren	15	1,68	46,67	53,33
-tzeko asmotan	1	0,11	100,00	0,00
-tzera	88	9,83	61,36	38,64

3.19 taula – Helburu-perpaus jokatuak eta ez-jokatuak corpusean

Helburu-perpaus jokatuak (34) sinplifikatzeko, sinplifikazio sintaktikoan bestelako ezabatzeak eta txertaketak egin ditugu. Mendeko perpausean, aditza nominalizatu (subjektua ezberdina delako) eta laguntzailea ezabatu dugu. Borondatea adierazteko, *nahi izan* aditza txertatu dugu. TE alternatibo gisa, *gura izan* proposatu dugu[32]. Perpaus nagusia bere horretan mantendu dugu. Esaldi berrien hurrenkera nagusia$_{jat}$-mendekoa$_{jat}$ da, corpuseko datuetan agertzen den bezala.

(34) a. Gorpua besarkatu eta muxukatu nuen, berriro bizitza eta zoriontasuna itzul **zedin**.

b. i. Gorpua besarkatu eta muxukatu nuen.

ii. Berriro bizitza eta zoriontasuna itzul*tzea nahi nuen*.

Helburu-perpaus ez-jokatuei dagokienez, lehenik eta behin aipatu behar da askoz usuagoak direla jokatuak baino, EGLUn adierazten den bezala eta gure corpus-azterketan, maiztasun-esperimentuan, konprobatu ahal izan dugun bezala[33]. Perpaus ez-jokatuak sinplifikatzeko (35), mendekoaren eta nagusiaren subjektua bera denez, ez dugu aditza nominalizatu. Horren ordez, partizipio bihurtu dugu aditza eta, jokatuetan bezala, *nahi izan* txertatu dugu.

(35) a. Abuztuaren amaieran beste goi bilera bat egitea aztertzen ari dira Israel eta PAN Palestinako Aginte Nazionala, Ekialde Erdiko bake prozesua susper**tzearren**.

b. i. Abuztuaren amaieran beste goi bilera bat egitea aztertzen ari dira Israel eta PAN Palestinako Aginte Nazionala.

ii. Ekialde Erdiko bake prozesua susper*tu nahi dute*.

[32] Maiztasunak: *nahi izan* 45.502; *gura izan* 848.

[33] Gogoan izan 3.4 taulan ikusi dugula jokatuak % 5,89an erabiltzen direla eta ez-jokatuak % 94,11n.

-tzeko asmotan, *-tzeko intentziotan*, *-tzeko intentzioarekin* eta antzeko egiturak ez sinplifikatzea erabaki dugu *asmo* eta *intentzio* hitzek jada argi uzten dutelako zein den helburua.

Maiztasunaren azterketan ikusi den bezala (3.19 taula), *-t(z)eko* gehien erabiltzen den egitura denez, ordezkapen sintaktikoen sinplifikazioan, gainontzeko helburu-perpaus ez-jokatuak horrekin ordezkatuz sinplifikatu ditugu. Corpusean aurkitu ez dugun *-tzekotan* egitura ere ordezkapenaren bitartez sinplifikatu dugu. Ordezka daitezkeen egiturak 3.20 taulan jaso ditugu.

Maiztasun gutxiko egiturak	**Ordezkapenerako hautagaia**
-tzekotzat; -tzeagatik; -tzearren; -tze alde(ra)	-tzeko
-tzekotan	-tzeko asmotan

3.20 taula – Maiztasun gutxiko helburu-perpausen egitura ez-jokatuak ordezkatzeko proposamenak

Helburu-perpausen berezitasuna da TEak ez direla adberbioak edo lokailuak, aditzak baizik.

Baldintza-perpausak

Baldintza-perpausek semantikoki baldintza bat, hipotesi bat adierazten dute perpaus nagusian azaltzen den ekintzarekiko. Mendeko perpausean, hipotesiak aurkezten dira, eta nagusian, gerta zitekeena edo daitekeena adierazten da. Hipotesiak gertakortasun-mailaren arabera definitzen dira eta bi motakoak dira: gertakorrak (orainaldian eta lehenaldian) eta gertagaitzak (alegialdian). Baldintza-perpausek perpaus adberbialen % 6,99 osatzen dute corpusean.

Baldintza-perpaus ez-jokatuetan gehien erabiltzen dena *-tuz gero((z)tik)* da, % 70,83ko maiztasunarekin. % 10eko muga *-tzekotan* egiturak gainditzen du, % 16,67 maiztasuna du. Baldintza-perpausek, hasierako esperimentuan ikusi dugun bezala, aditzaren aurretik joateko joera nabarmena dute. Egituraz egiturako analisia egitean ere joera bera mantentzen da jokatuetan eta gehien erabiltzen diren ez-jokatuetan (*-tuz gero((z)tik)* eta *-tzekotan*).

Baldintza-perpaus jokatuak sinplifikatzeko proposamenak gertagarritasunaren arabera emango ditugu. Lehenik, errealak direnak azalduko ditugu eta, ondoren, irrealak. Errealen barnean bi multzo bereiziko ditugu: adizkia orainaldian dutenak, eta adizkia lehenaldian dutenak.

Euskarazko egitura	Kopurua	Ehunekoa	Aurretik	Atzetik
ba-	249	50,00	81,12	18,88
-tuz gero((z)tik)	34	70,83	85,29	14,71
-tu ezkero(an/k/tino/((z)tik))	1	2,08	0,00	100,00
-tu ezean	2	4,17	50,00	50,00
-tzekotan	8	16,67	100,00	0,00
-tzera(t)	3	6,25	100,00	0,00

3.21 taula – Baldintza-perpaus jokatuak eta ez-jokatuak corpusean

Gertagarritasun erreala dutenak sinplifikatzeko, mendeko perpausaren hasieran *Demagun* txertatu dugu eta esaldi horretako aditzari *-ela* menderagailua gehitu diogu, perpaus konpletibo bat sortuz[34]. TE alternatiboa *Eman dezagun* da. Perpaus nagusian *Kasu horretan* txertatu dugu. Esaldi berrien hurrenkera mendekoa$_{jat}$-nagusia$_{jat}$ da, corpuseko datuetan ikusi dugun bezala, hurrenkera logikoa (baldintza-ondorio) betetzen baita. Orainaldiko erreal bat (36) adibidean ikus dezakegu eta lehenaldiko bat (37) adibidean.

(36) a. Partidua bide onetik **badoa**, Arestes eta Haritz Garcia gazteek minutu batzuk jokatzeko aukera izanen dute.

b. i. *Demagun* partidua bide onetik doa*la*.

ii. *Kasu horretan* Arestes eta Haritz Garcia gazteek minutu batzuk jokatzeko aukera izanen dute.

(37) a. Gizon hura armaturik baldin **bazegoen**, Salahadin prest zegoen tiro egiteko.

b. i. *Demagun* hura armaturik zegoe*la*.

ii. *Kasu horretan* Salahadin prest zegoen tiro egiteko.

Baldintza-perpaus irrealetan ekintza irreala da eta hori argi uztea da gure sinplifikazio-proposamenaren helburua. Horregatik, mendeko perpausaren polaritatea aldatu dugu; hau da, perpausa ezezkoan (38) badago, baiezkoan eman dugu, eta baiezkoan badago (39), ezekoan. Esaldi berriaren aditza orainaldian eman dugu. Perpaus nagusian *Bestela* txertatu dugu. Errealetan bezala, esaldi berrien hurrenkera mendekoa$_{jat}$-nagusia$_{jat}$ da.

[34]Orain arte mendeko perpausak ezabatzearen alde egin dugu, baina kasu honetan irtenbide hori egokiena eta eraginkorrena iruditu zaigu, *-ela*-k, hain zuzen, "erreal" izaera ematen diolako eta "demagun"-ek eskatzen duelako.

(38) a. Gorputzak ahal duena buruak galaraziko ez **balio**, Alex Zulle faborito sendoenetakoa litzateke, (..)

b. i. Gorputzak ahal duena buruak galarazten dio.

ii. *Bestela*, Alex Zulle faborito sendoenetakoa litzateke, (..)

(39) a. Komunikabideen alorrean olinpiar txapeldunik izendatu beharko **balitz**, Seven telebista kate publikoko The Dream-ek eramango luke urrea.

b. i. Komunikabideen alorrean olinpiar txapeldunik *ez* da izendatu behar.

ii. *Bestela*, Seven telebista kate publikoko The Dream-ek eramango luke urrea.

Txertatze-elementuen laburpena 3.22 taulan jaso dugu.

Multzoa	**Txertatze-elementuak**	**Txertatze-elementu alternatiboak**	**Hurrenkera**
Erreala (orainaldia)	Demagun$_{men}$ Kasu horretan$_{nag}$	Eman dezagun$_{men}$	mendekoa$_{jat}$-nagusia$_{jat}$
Erreala (lehenaldia)	Demagun$_{men}$ Kasu horretan$_{nag}$	Eman dezagun$_{men}$	mendekoa$_{jat}$-nagusia$_{jat}$
Irreala	(Polaritate aldaketa)$_{men}$ Bestela$_{nag}$	-	mendekoa$_{jat}$-nagusia$_{jat}$

3.22 taula – Baldintza-perpausen txertatze-elementuak

Aditz ez-jokatua duten perpausetan goikoaren antzeko sailkapena egitea eta sinplifikazio sintaktikoko proposamenak egitea zaila egin zaigu. Beraz, perpaus horiek sinplifikatzeko, ordezkapen sintaktikoen sinplifikazioa baino ez dugu proposatuko. Ordezkatzeko proposamenak 3.23 taulan ikus daitezke.

Maiztasun gutxiko egiturak	**Ordezkapenerako hautagaia**
-tuenenean; -tzera(t)	-tuz gero
-tu ezin	-tu ezean
-tzez gero; -tzekoz	-tzekotan

3.23 taula – Maiztasun gutxiko baldintza-perpausen egitura ez-jokatuak ordezkatzeko proposamenak

Baldintza-perpausen sinplifikazio-proposamenaren berezitasunak dira baldintza erreala adierazten duten perpaus jokatuetan perpaus konpletibo bat sortu dugula eta irrealetan, polaritate aldaketa egin dugula.

3.3.5 Aposizio-sintagmak

Aposizio-sintagmak izen-sintagma batekin batera azalpen bat eman edo karguaren (hondarkia) berri ematen duten izen-sintagmak dira. Euskaraz bi motako aposizio-sintagmak daude: 1) izen-sintagma baten barnean gertatzen direnak [1A mota, entitatea aurretik (40a) eta 1B mota, entitatea ondoren (40b)] eta 2) izen-sintagma osoak aposatuz egiten direnak (40c). Bi motak (40d) eta lehenengo motako bi hurrenkerak (40e) konbina daitezke.

(40) a. Luis Uranga presidenteak (...)
b. Errealeko presidente Luis Uranga (...)
c. Jakinduria hori, guretzat harrapezina dena, (...)
d. Simon Peres laborista, Israelgo lehen ministro izana, (...)
e. Vatikanoko Estatuekiko Harremanetarako idazkari Jean Louis Tauran artzapezpikuak (...)

Aposizio-sintagmak sinplifikatzeko (41) urrats hauek jarraitu ditugu:

1. Aposizio-sintagma bi izen-sintagmen mugan banatu: [Jasser Arafat]$_{s_1}$ [buru palestinarra]$_{s_2}$ [Egiptoko presidente]$_{s_3}$ [Hosni Mubarak-ekin]$_{s_4}$
2. Hondarkia perpaus nagusitik ezabatu: Jasser Arafat Hosni Mubarak-ekin bildu zen atzo Kairon.
3. Aposizio-sintagmekin esaldi berriak osatu
   3.1. Kasu-markarik badute, kasu-marka horiek ezabatu Hosni Mubarak-ekin -> Hosni Mubarak
   3.2. Izen-sintagma horiei absolutibo kasua gehitu, ez badute: Egiptoko presidente -> Egiptoko presidentea
   3.3. Aposizio-sintagma osatzen zuten bi izen-sintagmekin eta *izan* aditzaren 3. pertsona singularrarekin (*da*) esaldi berri bat osatu
4. Esaldien barne-hurrenkera: 'entitatea$_{subj}$ + hondarkia$_{pred}$ + *da*.'

5. Esaldi berriak ordenatu: nagusia$_{jat}$-apos$_{jat}$

6. Puntuazio-markak egokitu, eta akatsik badago zuzendu

Kontuan izan (41) adibidean bi aposizio daudela eta biak sinplifikatu ditugula.

(41) a. **Jasser Arafat buru palestinarra Egiptoko presidente Hosni Mubarak-ekin** bildu zen atzo Kairon.

b. i. Jasser Arafat Hosni Mubarak-ekin bildu zen atzo Kairon.

ii. Jasser Arafat buru palestinarra *da*.

iii. Hosni Mubarak Egiptoko presidentea *da*.

Izen-sintagma osoak aposatuz egiten diren aposizio-sintagmak (42) sinplifikatzeko ere, komak banatzen dituen bi izen-sintagmak banatu ditugu eta beste motarekin egin dugun bezala, bi osagaiekin esaldi bat osatu dugu.

(42) a. **Aitor Mendiluzek, hogei urteko andoaindar bertsolariak,** irabazi zuen Gipuzkoako txapelketa.

b. i. Aitor Mendiluzek irabazi zuen Gipuzkoako txapelketa.

ii. Aitor Mendiluze hogei urteko andoaindar bertsolaria *da*.

Aposizio-sintagmak sinplifikatzeko proposamenarekin esaldi eta sintagma laburragoak lortzeaz gain, esaldi berrietan entitatearen identifikazioa erraztu egiten dugu.

3.3.6 Egitura parentetikoak

Egitura parentetikoak testuan tarteki moduan (parentesien artean) integratuta dauden baina aldi berean nahiko independenteak diren modifikatzaileak dira. Testuari informazio osagarria ematen diote, eta sintagmez, perpaus batez edo esaldi batez osa daitezke. Kasu batzuetan egitura horiek ez daude esanahiari lotuta ez semantikoki ezta pragmatikoki ere, baina testuinguruagatik ulertzen dira (Dehé eta Kavalova, 2007). Beste kasu batzuetan, estiloaukeraren ondorio dira informazio biografikoa, esaterako. Atal honetan informazio biografikoa[35] azaltzen duten egitura parentetikoak, hain zuzen, izango ditugu aztergai.

[35]Informazio biografikoa diogunean oinarrizko informazio biografikoa adierazi nahi dugu, jaiotza- eta heriotza-datuak, alegia.

Informazio biografikoa duten egitura parentetikoak sinplifikatzeko, perpaus nagusiarekin eta tokien eta daten informazioarekin esaldi berriak sortu ditugu. Mota horretako egitura parentetikoak *Wikipedia*n maiz agertzen dira eta bertatik erauzitako adibideak landu ditugu.

Horrela, bada, hildako pertsona bati buruzko datu biografikoak (43) ditugunean tartekian, urrats hauek jarraitu ditugu:

1. Tartekiak esalditik banatu: [Ernest Rutherford, Nelsongo lehenengo baroia, fisika nuklearraren aita izan zen.]$_{p_1}$ [(Brightwater, Zeelanda Berria, 1871ko abuztuaren 30a - Cambridge, Ingalaterra, 1937ko urriaren 19a)]$_{t_1}$

2. Tartekiko jaiotza- eta hiltze-informazioa banatu: [Brightwater]$_{t_{1a1}}$ [Zeelanda Berria]$_{t_{1a2}}$ [1871ko abuztuaren 30a]$_{t_{1a3}}$ [Cambridge]$_{t_{1b1}}$ [Ingalaterra]$_{t_{1b2}}$ [1937ko urriaren 19a]$_{t_{1b3}}$

3. Tartekiko informazioarekin esaldi berriak osatu

   3.1. Pertsonaren izenarekin (perpaus nagusitik atera dugun informazioa) eta jaiotza-informazioarekin esaldi bat osatu, patroi honi jarraituz: 'pertsonaren izena (absolutiboan) + jaiotza-eguna (inesiboan) + koma + jaiotza-tokia (inesiboan) + *jaio zen*'

   3.2. Tartekian toki bat baino gehiago baldin badago, beste esaldi bat sortu dugu patroi honi jarraituz: 'tokia + bigarren tokia (inesiboan) + *dago*' Tartekian dauden toki guztiekin gauza bera egin dugu (toki-xehetasunak).

   3.3. Heriotza-datuak dituen esaldia osatu: 'pertsonaren izena (absolutiboan) + heriotza eguna (inesiboan) + koma + heriotza-tokia (inesiboan) + *hil zen*'.

   3.4. Heriotzaren datuetan toki bat baino gehiago agertzen badira, toki-xehetasunak dituzten esaldiak gehitu

4. Esaldiak testuan esaldiak ordenatu: 1) perpaus nagusia 2) jaiotza-datuekin sortutako esaldia 3) jaiotza-datuaren toki-xehetasunak (baldin badaude, eta behar adina esaldi) 4) heriotza-datuekin sortutako esaldia eta 5) heriotza-datuaren toki-xehetasunak (baldin badaude eta behar adina esaldi)

5. Jatorrizkoan akatsik egonez gero, zuzendu eta puntuazio-markak egokitu

(43) a. Ernest Rutherford, Nelsongo lehenengo baroia, **(Brightwater, Zeelanda Berria, 1871ko abuztuaren 30a - Cambridge, Ingalaterra, 1937ko urriaren 19a)** fisika nuklearraren aita izan zen.

b. i. Ernest Rutherford, Nelsongo lehenengo baroia, fisika nuklearraren aita izan zen.

ii. Ernest Rutherford 1871ko abuztuaren 30*ean,* Brightwater*en jaio zen.*

iii. Brightwater Zeelanda Berria*n dago.*

iv. Ernest Rutherford 1937ko urriaren 19*an,* Cambridge*n hil zen.*

v. Cambridge Ingalaterra*n dago.*

Pertsona bizirik baldin badago, berriz, hildakoekin bezala egin dugu, baina ez dugu heriotzari buruzko informaziorik eman, jakina.

3.4 Laburpena

Kapitulu honetan konplexutzat hartu ditugun fenomenoak eta horien eskuzko sinplifikazio-proposamenak aurkeztu ditugu. Aztergai izan ditugun fenomenoak koordinazioa, perpaus osagarriak, perpaus erlatiboak, perpaus adberbialak, aposizio-sintagmak, egitura parentetikoak eta zenbait postposizio-sintagma izan dira. Horretaz gain, sintaktikoki sinplifikatu ahal izateko, perpausek luzera minimoa bete behar dutela zehaztu dugu. Sinplifikazio sintaktikoko proposamenak, oro har, dira: i) esaldiak, sintagmak edo egitura parentetikoak banatu; ii) perpausetatik esaldi sinplifikatuak berreraiki ezabatzeak eta txertaketak eginez; iii) sortutako esaldi berriak testuan ordenatu eta iv) esaldiak zuzenak diren (ortografikoki eta ortotipografikoki) egiaztatu. Proposamen horien helburua da jatorrizko testuaren esanahiari eusten dioten eta aditz bakarra duten esaldi berriak sortzea. Ordezkapen sintaktikoen sinplifikazioan (perpaus adberbial ez-jokatuak lantzeko), aldiz, maiztasun gutxiko egiturak maiztasun altuagoa duten baliokideekin ordezkatu ditugu.

Horrekin batera, fenomenoei dagozkien sinplifikazio-proposamenen bereizgarritasunak laburbildu ditugu. Perpaus koordinatuak sinplifikatzeko, juntagailuak kendu ditugu, baina aurkaritzakoen edo hautakarien kasuan jatorrizko perpausean duten juntagailua bigarren koordinatuaren hasieran txertatu dugu. Bigarren perpaus koordinatuan aditza elidituta dagoenean, berriz, lehenengo koordinatutik berreskuratu dugu.

Perpaus osagarriak sinplifikatzeko, estilo-aldaketa egin dugu, eta aldaketa horretan kontuan izan dugu izenordainak eta aditzaren pertsonak alda daitezkeela. Puntuazio-markak ere estilo zuzenaren arabera eman ditugu. *-enez + aditz diskurtsiboak* egituraren eta landutako postposizio-sintagmen kasuan, txertaketak mendeko perpausean egin ditugu. Egitura horietatik sortutako esaldi sinplifikatuen hurrenkera, osagarrietan ez bezala, mendekoa$_{jat}$-nagusia$_{jat}$ eta post$_{jat}$-nagusia$_{jat}$ izatea erabaki dugu.

Perpaus erlatiboak (jokatuak zein ez-jokatuak) sinplifikatzeko, erlatibo eta aurrekari-mota (entitatea den ala izen arrunta den) kontuan izan ditugu. Aurrekariaren tratamendua eta esaldi berrien hurrenkera 3.24 taulan erakusten ditugu laburbilduta

Perpaus-mota	**Aurrekari-mota**	**Aurrekariaren tratamendua**	**Sinplifikatutako esaldien hurrenkera**
Erlatibo arrunta	Izen arrunta	Aurrekaria + erakuslea	mendekoa$_{jat}$-nagusia$_{jat}$
	Entitatea	Aurrekaria	mendekoa$_{jat}$-nagusia$_{jat}$
Zein erlatiboa	Izen arrunta	Aurrekaria + erakuslea	nagusia$_{jat}$-mendekoa$_{jat}$
	Entitatea	Aurrekaria	nagusia$_{jat}$-mendekoa$_{jat}$

3.24 taula – Perpaus erlatiboen sinplifikazio-proposamenen laburpena

Perpaus adberbialak sinplifikatzeko, bi sinplifikazio-mota proposatu ditugu: sinplifikazio sintaktikoa eta ordezkapen sintaktikoen sinplifikazioa. Sinplifikazio sintaktikoko proposamenetan, TE lehenetsiez gain, TE alternatiboak proposatu ditugu. Adberbial motaren edo azpimotaren arabera, esaldi berrien hurrenkera ezberdinak definitu ditugu. Perpaus adberbialetan txertatu behar diren elementuen eta sinplifikatutako esaldien hurrenkeraren laburpena jaso dugu 3.25 taulan. Aurretik jaso den kasuetan, zein taulatan ageri den adierazi dugu.

Perpaus-mota	**Txertatze-elementua**	**Txertatze-elementu alternatiboak**	**Hurrenkera**
Denbora	3.8 taula	3.8 taula	3.8 taula
Kausa	3.11 taula	3.11 taula	3.11 taula
Kontzesioa	Hala ere$_{nag}$	Nolanahi ere$_{nag}$; Edonola ere$_{nag}$; Hala eta guztiz ere$_{nag}$	mendekoa$_{jat}$-nagusia$_{jat}$

(Jarraipena hurrengo orrialdean)

Perpaus-mota	Txertatze-elementua	Txertatze-elementu alternatiboak	Hurrenkera
Modua	Hala$_{nag}$	Honela$_{nag}$; Horrela$_{nag}$; Modu horretan$_{nag}$; Era berean$_{nag}$; Era horretan$_{nag}$	mendekoa$_{jat}$-nagusia$_{jat}$
Ondorioa	Ondorioz$_{men}$	Beraz$_{men}$; Hortaz$_{men}$; Honenbestez$_{men}$	nagusia$_{jat}$-mendekoa$_{jat}$
Helburua	nahi izan$_{men}$	gura izan$_{men}$	nagusia$_{jat}$-mendekoa$_{jat}$
Baldintza	3.22 taula	3.22 taula	mendekoa$_{jat}$-nagusia$_{jat}$

3.25 taula – Perpaus adberbialen txertatze-elementuen, txertatze-elementu alternatiboen eta esaldien hurrenkeraren laburpena

Ordezkapen sintaktikoen sinplifikazioko proposamenak edo non ageri diren 3.26 taulan jasotzen dira.

Mota	Maiztasun gutxiko egiturak	Ordezkapenerako hautagaia
Denbora	3.9 taula	3.9 taula
Kausa	*-tzearren*	-tzegatik
Kontzesioa	*-tuagatik; -tuz gero ere, -tuta (gabe/ezta) ere; -tzearren; -ik ere*	-tu arren
Modua	3.16 taula	3.16 taula
Helburua	3.20 taula	3.20 taula
Baldintza	3.23 taula	3.23 taula

3.26 taula – Maiztasun gutxiko egiturak ordezkatzeko proposamenak

Aposizio-sintagmak eta egitura parentetikoak sinplifikatzeko, egitura horietan aurkitu ditugun elementuekin esaldi berriak osatu ditugu. Esaldi berri horien barne-hurrenkera azterketa honetan zehaztu dugu.

Kapitulu honetan aurkeztu ditugun sinplifikazio-proposamenak nola automatizatuko diren 4. kapituluan azalduko dugu.

4

Egitura konplexuen tratamendu automatikorantz

Kapitulu honetan, egitura konplexuak tratatzeko eta Euskarazko Testuen Sinplifikatzailea (EuTS) sistema sortzeko hartu ditugun erabakiak eta behar ditugun tresnak azalduko ditugu. Batetik, konplexutzat jo ditugun egitura sintaktikoen sinplifikazio-proposamenak nola automatizatu, noiz sinplifikatu eta norentzat sinplifikatu zehaztuko dugu. Eta, bestetik, testuen konplexutasuna automatikoki analizatzeko eta testuak automatikoki sinplifikatzeko beharrezkoak diren tresnak aurkeztuko ditugu.

4.1 Sarrera

Azterketa linguistikoan, egitura konplexutzat luzera minimo jakin bat eta aditz bat baino gehiago dituzten esaldiak (perpaus elkartuak), zenbait postposizio-egitura, aposizio-sintagmak eta egitura parentetikoak hartu ditugu. Azterketa horretan oinarrituta, sinplifikazio-proposamenak egin ditugu, eta kapitulu honetan sinplifikazio-proposamen horiek erregela bihurtuko ditugu. Horrekin batera, erregela horiek noiz eta zeren arabera aplikatu behar diren definituko dugu; hau da, sistemaren oinarri linguistikoak definituko ditugu eta testuak sinplifikatuko dituen sistema nolakoa izatea nahi dugun zehaztuko dugu hemen. Horretaz gain, erregela horiek automatizatzeko behar ditugun tresnak aurkeztuko ditugu.

Aurrekarietan (2. kapituluan) azaldu dugun bezala, testuak automatikoki sinplifikatzeko sistemek bi hurbilpen nagusi jarraitzen dituzte: datuetan oina-

rritutakoa eta ezagutza linguistikoan oinarritutakoa. Datuetan oinarritutako hurbilpenek datu-andana handiak behar dituzte, eta, ikusi dugun bezala, ingelesez *Wikipedia* eta *Simple Wikipedia* erabili izan dituzte. Frantsesez, haurrei zuzendutako *Vikidia* (haurren *Wikipedia* bezala har daitekeena) erabili dute; duela gutxi *Vikidia* euskaraz[1] sortu da. Baliabide horien arazoa[2] da, euskarazko *Vikidia*ren kasuan txikia izateaz gain, testuak ez direla sinplifikatuak; hau da, *Wikipedia* arrunteko artikuluaren testu sinplifikatua ez da *Simple Wikipedia*n aurkituko duguna, ez eta *Vikidia*n ere. Testu sinpleak biltzen dituzten baliabideetan aurkituko ditugun testuak gai berari buruzko artikuluak dira, baina bertsio sinplean; sinplifikazioa aztertzeko eta automatikoki ikasteko, berriz, testu paraleloak (sinplifikatuak) behar ditugula uste dugu. Ezagutza linguistikoan oinarritutako lanak, ordea, hizkuntzalariek edo adituek idatzitako erregeletan oinarritzen dira. Guk azken hori aukeratu dugu, sinplifikazio-lanetan hasi ginenean bestelako baliabiderik ez genuelako eta baliabide urriko hizkuntzetarako egokia iruditu zaigulako. Beraz, EuTS sistema erregeletan oinarrituko da.

4.2 Konplexutasuna tratatzeko erabakiak

EuTS sistemaren diseinuan hurbilpenaz eta sinplifikazio-motaz gain, konplexutasuna tratatzeko ere kontuan hartu behar da noiz eta norentzat sinplifikatuko den. Noiz eta zeinentzat sinplifikatu behar dugun jakiteko, sinplifikazio-erabakien algoritmoa aurkeztuko dugu, eta testuak zeinentzat sinplifikatuko ditugun zehazteko, sinplifikazio-mailak aurkeztuko ditugu. Sinplifikazio-motak nola gauzatuko diren ere azalduko dugu.

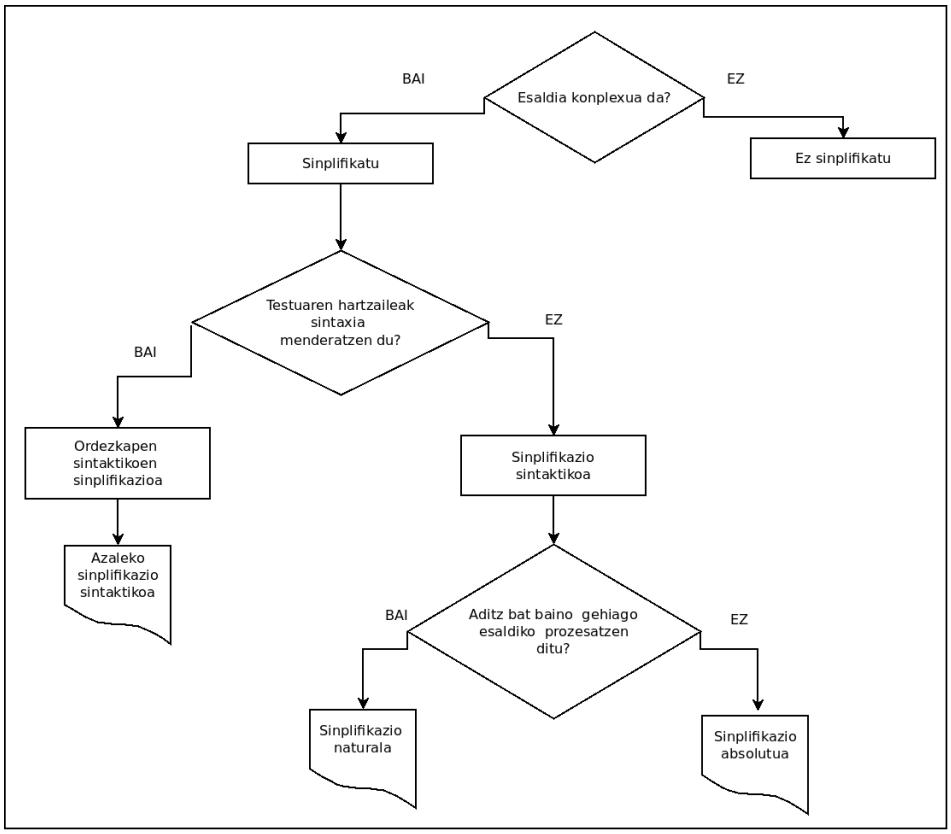
4.2.1 Sinplifikazio-erabakien algoritmoa

Noiz eta zeinentzat sinplifikatu nahi dugun aukeratzeko sinplifikazio-erabakien algoritmoa (4.1 irudia) definitu dugu. Proposatu dugun sinplifikazio-erabakien algoritmoan bi faktorek hartzen dute parte: i) jatorrizko testuaren konplexutasuna eta ii) testuaren helburu edo hartzaile izango den pertsona edo HPko aplikazioa. Horien arabera, testua nola sinplifikatuko den

[1] 2015eko ekainean sortu da eta https://eu.vikidia.org/wiki/Azala atarian kontsulta daiteke (2015eko urrian atzituta).

[2] Datu gabezia eta datuen egokitasuna TSAn arazo bat da, oro har, ez da soilik baliabide urriko hizkuntzen arazoa.

(sinplifikazio-mota) eta zein sinplifikazio-mailatara egokituko den erabakiko da.



4.1 irudia – Sinplifikazio-erabakien algoritmoaren errepresentazio grafikoa

Algoritmo horren arabera, testu bat edo esaldi bat dugula, erabaki beharreko lehendabiziko gauza da ea sarrera hori konplexua edo sinplea den. Hori erabakitzeko, testuak sinpleak ala konplexuak diren adierazten duen Erre-Xail sistema (ikus 5. kapitulua) erabiliko du. Sarrera-testu hori ez bada konplexua, ez du sinplifikatuko; konplexua bada, berriz, helburu-taldearen edo testuaren hartzailearen beharrak izango ditu kontuan, gizakien kasuan sintaxia menderatzen duen (B2 mailatik gorako ezagutza duen) eta, HPko tresnen kasuan, entrenamendu-corpusak eta tresnek zein ezaugarri dituzten, esaldi luzeak edo egitura dialektalak esaterako.

Testuaren hartzaileak sintaxia menderatzen badu, hau da, B2 mailatik go-

rako ezagutzak baditu, ordezkapen sintaktikoen sinplifikazioa da egingo duen sinplifikazio-mota eta testua azaleko sinplifikazio sintaktikoko sinplifikazio-mailara egokituko du. Testuaren hartzaileak sintaxia menderatzen ez badu, hau da, B2 mailara ez bada iristen, sinplifikazio sintaktikoa egingo du eta sinplifikazio sintaktikoaren prozesua aplikatuko du.

Sinplifikazio sintaktikoa aplikatzen duenean, beste bi sinplifikazio-maila bereiziko ditu berriro ere, hartzailearen mailaren arabera: sinplifikazio naturala eta sinplifikazio absolutua. Maila horiek bereizteko, testuaren hartzaileak esaldi bakoitzean aditz bat baino gehiago prozesatzen dituen kontuan izango du. Sinplifikazio-maila horiek 4.2.2 atalean definitu eta azalduko ditugu.

Testuen konplexutasuna analizatzea testuak sinplifikatu behar diren jakiteko, euskaraz ez ezik ingelesez, Brasilgo portugesez, italieraz eta alemanez ere egin da (Feng *et al.*, 2010; Aluísio *et al.*, 2010; Dell'Orletta *et al.*, 2011; Hancke *et al.*, 2012; Vajjala eta Meurers, 2014a). Analisi horrekin konplexuak diren testuak identifikatzen dira, zein sinplifikatu behar den jakiteko.

Zein sinplifikazio-mailatan sinplifikatu behar den erabakitzeko, Brasilgo portugesaren kasuan ikasketa automatikoa erabili dute; sinplifikazio-erabaki horiek eskuz sinplifikatutako testuetatik ikasi ditu sistemak (Gasperin *et al.*, 2009a). Esaldiak noiz banatu behar diren (hau da, sinplifikazio sintaktikoa litzatekeena noiz aplikatu behar den) jakiteko, sinplifikazio naturalaren kasuan zein esaldi banatu behar diren jakiteko, sailkatzaile bitar bat erabili dute Brasilgo portugesaren kasuan (Gasperin *et al.*, 2009b). Ingelesaren kasuan, berriz, esaldi konplexuaren luzeran oinarritu dira (Zhu *et al.*, 2010), eta gaztelaniaren kasuan ikasketa automatikoan (Štajner *et al.*, 2013). Gaztelaniaren kasuan, zein fenomeno sinplifikatu behar diren jakiteko gramatikak erabili dituzte (Saggion *et al.*, 2015b).

4.2.2 Sinplifikazio-mailak

Atal honetan euskarazko testuak sinplifikatzeko bereizten ditugun sinplifikazio-mailak aurkeztuko ditugu. Hiru dira oraingoz definitu ditugun sinplifikazio-mailak: azaleko sinplifikazio sintaktikoa, sinplifikazio naturala, eta sinplifikazio absolutua. Sinplikazio-maila horiek, esan bezala, testuaren hartzailearen euskara-maila edo HPko tresnaren beharrak hartzen dituzte kontuan; beraz, lantzen diren fenomenoak berak dira, baina fenomeno horien aplikazioa aldatu egiten da hartzailearen arabera. Definitutako hiru sinplifikazio-mailak hauek dira eta aurrerago (44) adibidean ikusiko ditugu:

- **Azaleko sinplifikazio sintaktikoa (ASS):** Azaleko sinplifikazio sintaktikoa, ingelesez *shallow syntactic simplification* (SSS), maiztasunetan oinarritutako egitura baliokideen ordezkapena da. Maila honetan perpaus adberbial ez-jokatuak landuko dira. Sinplifikazio-maila honetan testua ulergarriagoa emango da jatorrizko sintaxiaren konplexutasunari uko egin gabe, hau da, egitura ezagunagoak emanez sinplifikatu nahi dugu testua. ASSa egitura-aldaketarik behar ez duten helburu-taldeentzat egokia izango litzateke, hau da, sintaxia menderatzen, baina egitura guztiak ezagutzen ez dituztenentzat (maila aurreratua, B2tik aurrera). HPko sistemei edo tresnei dagokienez, ASSa egokia da entrenamendu-corpusetan maiztasun gutxiko egiturak izan ez dituztenentzat.

- **Sinplifikazio naturala (SN):** Sinplifikazio naturala, ingelesez *natural simplification* (NS), testu baten konplexutasuna gutxitzea da esaldi laburragoak eginez, baina luzeraren aldetik oraindik "natural" izan daitezkeen esaldiak sortuz, hau da, ez oso laburrak ez eta oso sinpleak ere diren esaldiak sortuz. SNean sinplifikazio sintaktikoa gauzatuko da, eta perpaus koordinatuak, mendeko perpaus jokatuak, aposizio-sintagmak eta egitura parentetikoak izango dira tratatuko diren fenomenoak. Sinplifikazio-mota hori aurrekoa baino gogorrago den heinean (sintaxian egitura-aldaketak egiten baitira), euskara-maila baxuagoa dutenentzat (erdi-maila, B1-B2) edo esaldi laburrak hobeto prozesatzen dituzten tresnentzat bideratuta dago.

- **Sinplifikazio absolutua (SA):** Sinplifikazio absolutuan, ingelesez *absolute simplification* (AS), testuan dauden egitura konplexu guztiak sinplifikatuko dira: koordinazioa, mendeko perpaus jokatu eta ez-jokatuak, aposizio-sintagmak, egitura parentetikoak eta adierazpenak adierazten dituzten postposizio-sintagmak[3]. Sinplifikazio-mota honetan testua bere bertsio sinpleenera eramaten da. SAtik sortu diren testuak euskara ikasten hasi direnentzat (hastapen-maila, A1-A2) eta oso esaldi laburretan informazioa erauzi behar duten aplikazioentzat dira egokiak.

Azaldu ditugun sinplifikazio-maila horiek (44a) jatorrizko esaldiari aplikatuko dizkiogu. (44b-i) esaldia ASS mailara egokitu dugu eta, horregatik,

[3]Gainontzeko postposizio-sintagmak ez ditugu sistematikoki aztertu, perpaus mailako sintaxian kontzentratu garelako oro har.

soilik maiztasun gutxiko egitura sintaktikoak (helburu-perpaus ez-jokatua) ordezkatu ditugu, sintaxian egitura-aldaketarik egin gabe. (44b-ii) esaldia (44a) esaldiari SNa aplikatzearen emaitza da, eta ASSa egiteaz gain (helburu-perpaus ez-jokatua), perpaus jokatuak dituzten esaldietan sinplifikazio sintaktikoa aplikatu dugu; horren ondorioz, perpaus koordinatuak sinplifikatu ditugu *igo* eta *jaitsi* aditzak dituzten perpausak banatuta eta *eta* juntagailua ezabatuta. (44b-iii) esaldia SAra egokitu dugu, eta esaldian agertzen diren egitura guztiak (perpaus koordinatuak, helburu-perpaus ez-jokatua eta denbora perpaus ez-jokatua) sinplifikatu ditugu. (44b-ii) eta (44b-iii) esaldiak sortzeko sintaxian egitura-aldaketa egin ditugu.

(44) a. 1991 eta 1993an errepideko Munduko Txapelketak eta bi Tour irabazi **ondoren**, mendian gora aise igo**tzearren** pisua galtzen hasi zen, **eta** 1994. urtean 48 kiloko infernura jaitsi zen, anorexiara.

b. i. 1991 eta 1993an errepideko Munduko Txapelketak eta bi Tour irabazi *ondoren*, mendian gora aise igo*tzeko* pisua galtzen hasi zen, eta 1994. urtean 48 kiloko infernura jaitsi zen, anorexiara.

ii. 1991 eta 1993an errepideko Munduko Txapelketak eta bi Tour irabazi *ondoren*, mendian gora aise igo*tzeko* pisua galtzen hasi zen. 1994. urtean 48 kiloko infernura jaitsi zen, anorexiara.

iii. 1991 eta 1993an errepideko Munduko Txapelketak eta bi Tour irabazi zituen. *Ondoren*, pisua galtzen hasi zen. Mendian gora aise igo *nahi zuen*. 1994. urtean 48 kiloko infernura jaitsi zen, anorexiara.

Sinplifikazio-mailak Brasilgo portugesaren kasuan ere bereizi dituzte. Maila horiek sinplifikazio naturala (*natural simplification*) eta sinplifikazio absolutua (*strong simplification*) dira, eta jendearen alfabetatze-mailaren araberakoak dira. Lehenengoa oinarrizko maila (*basic*) dutenei zuzenduta dago, eta bigarrena, berriz, oso oinarrizkoa (*rudimentary*) dutenei. Bien arteko diferentzia aplikatzen diren eragiketetan datza. Sinplifikazio naturalean esaldien banaketa eta esaldien hurrenkera-aldaketak zuhurtasunez aplikatzen dituzte, eta absolutuan, aldiz, eskuliburu batean definitutako eragiketak aplikatzen dituzte. Dena dela, bi mailetan esaldia hurrenkera kanonikora pasatzen dute eta ahots pasiboa aktibo bihurtzen dute (Gasperin *et al.*, 2009a).

4.2.3 Sinplifikazio-motak

Testuak automatikoki sinplifikatu diren lanetan, oro har, sinplifikazio lexikala eta sintaktikoa egin dira. Lan batzuetan, gainera, informazio gehigarria emanez egindako sinplifikazioak proposatu dira (ikus 2. kapitulua). Gure azterketa linguistikoan esan dugun bezala, sintaxian egitura-aldaketak egiten dituzten eta egitura-aldaketak egiten ez dituzten sinplifikazio-motak proposatu ditugu. Beraz, EuTS sistemak bi sinplifikazio-mota egingo ditu: i) sinplifikazio sintaktikoa eta ii) ordezkapen sintaktikoen sinplifikazioa. Sinplifikazio sintaktikoan egitura-aldaketak egingo ditugu, eta ordezkapen sintaktikoen sinplifikazioan, maiztasun gutxiko egiturak maiztasun altuagoa dutenekin ordezkatuko ditugu, egitura-aldaketarik egin gabe. Tesi-lan honetan egin dugun azterketa sintaxian oinarritzen denez, sinplifikazio lexikala etorkizuneko lanetarako utzi dugu.

Ordezkapen sintaktikoen sinplifikazioa

Ordezkapen sintaktikoen sinplifikazioak sinplifikazio lexikalaren teknikak erabiltzen ditu, baina sintaxian aplikatzen da. Sinplifikazio lexikalaren teknikak erabiltzen dituela aipatzen dugunean, esan nahi dugu egitura bat esanahi bera duen eta maiztasun altuagoa duen egitura batekin ordezkatzen duela eta ordezkapenak egiteko maiztasunetan oinarritzen dela.

Ordezkapen sintaktikoen sinplifikazioa azaleko ordezkapen sintaktikoak (*shallow syntactic substitutions*) eragiketaren bitartez aplikatuko du. Ordezkapenen erregelak egiteko baliokidetzak 3. kapituluan aurkeztu ditugu eta perpaus adberbial ez-jokatuentzat aztertu ditugu.

Sinplifikazio sintaktikoa

Azterketa linguistikoan ikusi dugun bezala (3. kapitulua), esaldiaren sintaxian egitura-aldaketak egiten dituzten sinplifikazio-proposamenak automatizatzeko erabakiak azalduko ditugu hemen. Automatizazio hori sinplifikazio sintaktikoaren prozesuaren bitartez egingo dugu. Hau da, sintaxian egitura-aldaketak egiten dituzten sinplifikazio-proposamenak sinplifikazio-erregela bihurtuko ditugu sinplifikazio sintaktikoaren prozesuaren bidez.

Sinplifikazio sintaktikoaren prozesuak 4 eragiketa ditu: i) banaketa, ii) esaldien berreraikitzea, iii) esaldien ordenatzea eta iv) esaldien zuzenketa eta egokitzapena. Eragiketa horiek hobeki ulertzeko, (45) adibidea izango

dugu aztergai, eta jatorrizko esalditik (45a) sinplifikatutako (45b-i) eta (45b-ii) esaldiak nola lortuko diren azalduko dugu.

(45) a. Asperren kasua emeki-emeki aitzinatu **ba**da **ere**, Sa Pintoren etorkizuna fite argituko da.

b. i. Asperren kasua emeki-emeki aitzinatu da.

ii. *Hala ere*, Sa Pintoren etorkizuna fite argituko da.

Jarraian, eragiketa horiek zehatzago azalduko ditugu:

1. **Banaketa (***splitting***)**: Eragiketa honetan esaldian dauden perpaus guztiak banatuko dira, aditzaz gain gutxienez beste bi osagarri edo adjuntu badituzte (luzera minimoa, ikus 4.2.3 azpiatala). Hau da, esaldi batek dituen perpaus guztiak identifikatuko dira, eta esalditik banatuko dira luzera minimoa badute. Honela, esaldian dagoen perpaus bakoitzarekin esaldi berri bat sortzeko hautagaia lortuko da. Perpausak banatzeaz gain, azterketa linguistikoan landutako adierazpenak adierazten dituzten postposizio-sintagmak, aposizio-sintagmak eta egitura parentetikoak ere banatuko dira.

   (45a) esaldian bi perpaus aurkitu eta banatu ditugu: mendeko perpaus adberbiala (kontzesio-perpausa) *Asperren kasua emeki-emeki aitzinatu bada ere,* eta perpaus nagusia *Sa Pintoren etorkizuna fite argituko da.*

2. **Esaldien berreraikitzea (***reconstruction***)**: Banatu diren perpausetatik esaldi sinpleak sortzea da eragiketa honen helburua. Bi azpieragiketa hartuko ditu bere gain:

   - **Ezabatzea (***removing***)**: Ezabatzen diren elementuak erlazio-markak (menderagailuak, kasu-markak, postposizioak eta hitz funtzionalak) dira. Euskara hizkuntza eranskaria denez, hemen inplementatuko diren erregelak (B eranskina), oro har, morfologian oinarrituko dira. Ezabatu behar diren erlazio-markak Erlazio_Marken_Zerrenda zerrendan (B eranskina) gordeko dira.
   - **Txertatzea (***adding***)**: Txertatzen diren elementuak adberbioak eta izen-sintagmak dira, oro har. Erregela jakin batzuetan, helburu-perpausen erregeletan esaterako, aditzak eta ezaugarri morfologikoak (morfemak, aspektu-markak eta aditz oinen markak) txertatuko dira. Elementu horien helburua esaldiak banatu ondoren,

ezabatzean galdu dituzten erlazioak berreskuratzea da, esaldi berriek jatorrizko esanahiari euts diezaioten. Hemen erabakiko da ere txertatze-elementu lehenetsia edo txertatze-elementu alternatiboa erabili behar den aurrerago ikusiko dugun bezala. TE lehenetsiak eta TE alternatiboak Txertatze_Elementuen_Zerrenda zerrenden zerrendan gordeko dira. TE alternatiboak dagokien zerrendaren bigarren posiziotik aurrera agertzen dira (B eranskina).

Adibidera itzuliz eta kontzesio-perpausen erregelari jarraikiz, mendeko perpausetik *ba- ere* erlazio-marka ezabatu dugu, horrela adierazten baita Erlazio_Marken_Zerrenda zerrendan: *Asperren kasua emeki-emeki aitzinatu da.* Perpaus nagusian, berriz, *Hala ere* lokailua txertatu dugu Txertatze_Elementuen_Zerrenda zerrenda kontzesio-perpausekin lotuta dagoelako eta maiztasun altuena duelako. Jatorrizko esaldian mendeko perpausaren ondoren koma dagoenez, TEaren ondoren ere koma[4] mantendu dugu: *Hala ere, Sa Pintoren etorkizuna fite argituko da.*

3. **Esaldien ordenatzea** (*reordering*): Eragiketa honek bi hurrenkeramaila ezberdinetan hartzen du parte:

   - Esaldi berrien barneko osagaien hurrenkera: Hurrenkera kanonikoa edo kognitiboki prozesatzeko kostu gutxien duten hurrenkerak erabil daitezkeen arren, oraingoz jatorrizko esaldiak duen hurrenkerari helduko zaio. Kognitiboki prozesatzeko kostu gutxien duten hurrenkerak ezagutzen ditugunean, horiek aplikatuko ditugu. Neurohizkuntzalaritzako lanen berri izan ahala, hurrenkera-erregela horiek jasoko ditugu. Orain arte ikusi ditugun lanen arabera, euskaraz SOV ordena errazagoa da prozesatzen OSVa baino (Erdocia *et al.*, 2009), baina gure sistemaren erregelak definitzeko estaldura handiagoko lanak beharko genituzke. Izan ere, lan horiek osagaien hurrenkera dute aztergai, eta guk adjuntuen kokapena ere beharko genuke. Dena den, aposizioetatik eta egitura parentetikoetatik sortzen diren esaldi berrien barne-hurrenkeraren ordenatzea hemen aplikatuko da.
   - Esaldi berrien arteko hurrenkera: Perpaus adberbialen kokapenaren azterketaren (ikus 3.3.4 atala) emaitzak eta hurrenkera logi-

[4]Kasu honetan TEa lokailua denez, koma jartzea beharrezkoa da.

koa erabiliko dira; esate baterako, baldintza eta kausa ondorioaren aurretik emango dira. Definitutako hurrenkera horiek Hurrenkeren_Zerrenda zerrendan gordeko dira.

Adibidean, esaldi berrien barneko hurrenkera dagoen bezala utzi dugu. Esaldi berrien hurrenkera, aldiz, kontzesio-perpaus jokatuei dagokien erregelan zehazten den eta Hurrenkeren_Zerrenda zerrendan jasotzen den bezala, esaldi berrien hurrenkera mendekoa$_{jat}$-nagusia$_{jat}$ da. Beraz, testuan lehendabizi *Asperren kasua emeki-emeki aitzinatu da* eman dugu, eta ondoren *Hala ere, Sa Pintoren etorkizuna fite argituko da.*

4. **Esaldien zuzenketa eta egokitzapena (***correction and adequation***)**: Sortu diren esaldi berrien zuzentasun gramatikala berrikustea eta kohesioa bermatzea da azken eragiketa honen helburua. Izan ere, sortzen diren testuak akastunak badira, testuaren ulermena eta prozesamendua erraztu baino gehiago zaildu egingo lirateke. Esaldi berrien puntuazio-markak ere hemen txukunduko dira.

   Adibideko lehen esaldiari puntua ipini diogu, eta, beste akatsik zuzendu behar ez denez, honako hau dugu prozesuaren emaitza: *Asperren kasua emeki-emeki aitzinatu da. Hala ere, Sa Pintoren etorkizuna fite argituko da.*

Sinplifikazio sintaktikoko prozesuko eragiketa bakoitza 6. kapituluan aurkeztuko dugun EuTS sistemaren arkitekturaren moduluekin dago lotuta; hau da, arkitekturaren modulu bakoitzak hemen aurkeztuko dugun eragiketa bat egingo du. Modulu horiek erabiliko duten informazio linguistikoa gramatiketan (banaketa eta zuzenketa eragiketetan) eta zerrendetan (berreraikitzea eta ordenatzea eragiketetan) kodetuko da. Sinplifikazio sintaktikoko prozesua algoritmo baten bitartez definituta eman dugu 4.2 irudian. Bertan, adibide gisa, mendeko perpaus adberbialen erregelen urratsak eragiketekin lotu ditugu eta zein baliabide behar diren adierazi dugu. Sinplifikazio sintaktikoaren erregelak B eranskinean jaso ditugu.

Sinplifikazio sintaktikoa egiteko definitu ditugun eragiketak beste hizkuntzetan ere aurkitu ditugu (ikus 2.3 taula). Lan horietan banaketa izendatzeko *splitting* edo *split* erabili izan dute. Esaldien berreraikitzeko eragiketari beste hizkuntzetan *transformation* esan diote. “Transformazio” hitzak guretzat, aldiz, zentzu zabalagoa dauka, eta, eragiketa honetan guk esaldiak berreraikitzen ditugunez, *reconstruction* edo “berreraikitze” erabiltzea erabaki dugu.

```
- Mend:  mendeko perpausaren mota
- Ald:  zenbat alditan egin den aldaketa mendeko perpaus honetan
hasiera

   1. Banatu_Esaldia_Perpausetan - Mugak

   2. Ezabatu_Erl_Marka_Mendeko_Perpausetik
      (Mend, Erlazio_Marken_Zerrenda)

   3. Aukeratu_Elementua_Txertatzeko
      (Mend, Ald, Txertatze_Elementuen_Zerrenda)

   4. Ordenatu (Mend, Hurrenkeren_Zerrenda)

   5. Zuzendu_PM_eta_Akatsak - Xuxen

bukaera
```

4.2 irudia – Perpaus adberbialak sintaktikoki sinplifikatzeko prozesuaren algoritmoaren errepresentazioa

Ezabatze-eragiketan beste hizkuntzetan hitz osoak dira ezabatzen direnak, eta horregatik egile batzuek *dropping* (erortzea, erortzen uztea) edo *delete* (ezabatu) esan diote eragiketa horri. *Delete* kontzeptuarekin, zehazki, hitz (lexikalak nahiz funtzionalak), sintagma, perpaus edo esaldi osoak (unitate bezala) ezabatzea adierazi ohi dute. Txertaketa izendatzeko, berriz, *insert* (txertatu) izena erabili dute. Esaldien ordenatzea izendatzeko, oro har, *reordering, reorder* edo *clause order* edo *ordering* erabili izan dute. Esaldien zuzenketa eta egokitzapena, berriz, ez dute hizkuntza guztietan kontuan hartu, eta sinplifikazio-prozesutik at mantendu dute. Hala ere, egokitzapenari *regeneration* deitu izan diote. Dena den, zuzenketa gure ustez ezinbestekoa da, sortuko ditugun testuen zuzentasuna (eta horrekin batera kohesioa) bermatzeko, eta behintzat jatorrizko testuek izan ditzaketen balizko akatsak ez ditugu barreiatuko; izan ere, zuzenketaren beharra analisi linguistikoan ikusi dugu. Termino horiek guztiak 4.1. taulan jaso ditugu.

Euskarazko terminoa	**Ingelesezko terminoa**	**Beste lanetan emandako terminoak**
Banaketa	*splitting*	*splitting*, *split*
Esaldien berreraikitzea	*reconstruction*	*transformation*
Ezabatzea	*removing*	*dropping*, *delete*

(Jarraipena hurrengo orrialdean)

Euskarazko terminoa	**Ingelesezko terminoa**	**Beste lanetan emandako terminoak**
Txertatzea	*adding*	*insert*
Esaldien ordenatzea	*reordering*	*reordering*, *reorder*, *clause order*, *clause ordering*
Esaldien zuzenketa eta egokitzapena	*correction and adequation*	*regeneration*

4.1 taula – Sinplifikazio-eragiketen terminoak

EuTS sistema garatzeko, noiz eta zeinentzat (sinplifikazio-erabakien algoritmoa), eta nola (sinplifikazio sintaktikoaren prozesua) erabakitzeaz gain, beste hiru erabaki nagusi hartu ditugu: i) esaldiak sinplifikatuak izan daitezen perpausek izan behar duten luzera minimoa, ii) txertatze-elementu alternatiboen erabilera eta iii) erregelak aplikatzeko hurrenkerak. Atal honetan erabaki horiek zergatik hartu ditugun azalduko dugu. Horretaz gain, etorkizunean integratu nahi ditugun ezaugarriak aurkeztuko ditugu.

Luzera minimoa. Sinplifikazio-proposamenetan ezarri dugun luzera minimoa perpaus bakoitzean (mendekoan zein nagusian) aditzaz gain bi argumentu edo adjuntu izatea da. Erabaki horrekin oso esaldi laburren sorrera eta txertatze-elementuen gehiegizko erabilera ekidin nahi dugu.

Horren adibidea da (46)ko esaldia. Sinplifikazio sintaktikoa gauzatuko bagenu, (46a) esaldiko mendeko perpausean aditzaz gain argumentu bakarra legoke (*hori* objektu zuzena) eta perpaus nagusian subjektua (*Josu Urrutia*) eta *ez* partikula. (46b-i) eta (46b-ii) esaldi sinplifikatuak irakurtzean, bien artean sortzen den etenak ulermena zailtzen duela iruditzen zaigu. Beraz, luzera minimoa betetzen ez bada, ez ditugu sinplifikazio sintaktikoaren erregelak aplikatuko.

(46) a. Hori galtzen du**enean**, ez da Josu Urrutia izango.
b. i. Hori galtzen du.
ii. Orduan, ez da Josu Urrutia izango.

Esaldien luzera kontrolatuko dugu erregelak aplikatzen goazen heinean. Beraz, erregelek murriztapen hori betetzen duten bitartean (47) aplikatuko ditugu, eta luzera minimoa betetzen ez denean erregelak aplikatzeari utziko diogu. Horregatik, sinplifikazio absolutua egin arren, (47a) esaldian *aldatzeko* eta *irabazteko* aditzek sortzen dituzten perpausei ez dizkiegu dagozkien erregelak aplikatu.

(47) a. Egoera alda**tzeko** garaipen bat behar dugu, **eta**, zelaia zalez gainezka egongo de**nez**, irabaz**teko** giro ezin hobea izango dugu.

b. i. Egoera aldatzeko garaipen bat behar dugu.
ii. Zelaia zalez gainezka egongo da.
iii. Horregatik, irabazteko giro ezin hobea izango dugu.

Erregelak aplikatzeko luzera-murriztapena esaldiak noiz banatu behar diren jakiteko ere erabili dute (Zhu *et al.*, 2010).

Txertatze-elementu alternatiboak. Txertatze-elementu alternatiboak proposatu ditugu txertatze-elementuak sor dezakeen monotoniarekin hausteko eta hitz bera oso gertu ez erabiltzeko. Beraz, TEa aukeratu baino lehen, lehenetsia jatorrizko esaldian edo aurreko eta ondorengo bietan ba ote dagoen egiaztatuko dugu. Testua sinplifikatzean behin baino gehiagotan erabili dugun edo testuan erregela hori aplikatu den ere izango dugu kontuan. Honekin testuaren kohesioa eta aberastasuna bermatu nahi ditugu.

(48) adibidean TE alternatiboa gehitu dugu erregela aplikatzean. Izan ere, kausa-perpausa sinplifikatzeko erregela aplikatzean ikusi dugu erregela horretako *Horregatik* TEa jatorrizko esaldian jada erabilita dagoela. Beraz, horren ordez *Hori dela-eta* alternatiboa txertatu dugu (48b-iii) esaldian.

(48) a. Horrez gainera, Bruselak dio laguntza haien berri ez zuela jakin bere garaian, Madrilgo Gobernuak jakinarazi ez zio**lako**, eta **horregatik** ere arauz kontrakoak direla.

b. i. Horrez gainera, Bruselak honako hau dio:
ii. "Madrilgo gobernuak ez digu jakinarazi.
iii. *Hori dela-eta*, laguntza haien berri ez dugu jakin bere garaian."
iv. Horregatik ere honako hau dio:
v. "Arauz kontrakoak dira."

Aztertu ditugun beste hizkuntzen lanetan ez dute txertatze-elementu alternatiborik proposatu.

Sinplifikazio-erregelen aplikazio-hurrenkera. Sinplifikazio sintaktikoaren prozesua aurrera eramateko hartu beharreko erabakia da erregelak zein hurrenkeratan aplikatu behar diren jakitea. Esaldian fenomenoak agertu

ahala sinplifikatu behar da (ezkerretik eskuinera)? Amaieran agertzen diren fenomenoekin hasi behar da (eskuinetik ezkerrera)? Hori dela-eta, fenomeno guztiak sinplifikatu behar direnean, esaldiaren mendekotasun-zuhaitzaren araberako erregelen aplikazioa proposatu dugu. Hau da, esaldiaren dependentzia-zuhaitza egingo dugu eta goitik behera hasiko gara erregelak aplikatzen. Erabakia zergatik hartu dugun (49) adibidearekin azalduko dugu.

Bi hurrenkera ezberdinetan sinplifikatu dugu (49a) esaldia, biak SN mailan. (49b) multzoko esaldietan fenomenoak agertu ahala (ezkerretik eskuinera) sinplifikatu ditugu eta (49c) multzoan, berriz, dependentzia-zuhaitzaren bertikaltasuna erabili dugu.

Jatorrizko esaldian dugun lehenengo fenomenoa helburu-perpaus adberbiala (*izan dadin*) da, eta perpaus hori sinplifikatzeko dagokion erregela[5] aplikatu dugu (49b-i) esaldia lortzeko. Ondoren perpaus konpletiboa (*eman behar dietela*) dugu, eta horren erregela[6] aplikatu dugu (49b-ii) esaldia lortzeko.

(49) a. akordioa posible **izan dadin**, sozialistek PPk egindako zazpi galderei baiezko erantzuna **eman behar dietela** esan zuen.

b. i. Sozialistek PPk egindako zazpi galderei baiezko erantzuna eman behar dietela esan zuen. Akordioa posible izatea nahi dute.

ii. Honako hau esan zuen: “Sozialistek PPk egindako zazpi galderei baiezko erantzuna eman behar diete.” Akordioa posible izatea nahi dute.

c. i. Honako hau esan zuen: “Akordioa posible izan dadin, sozialistek PPk egindako zazpi galderei baiezko erantzuna eman behar diete.”

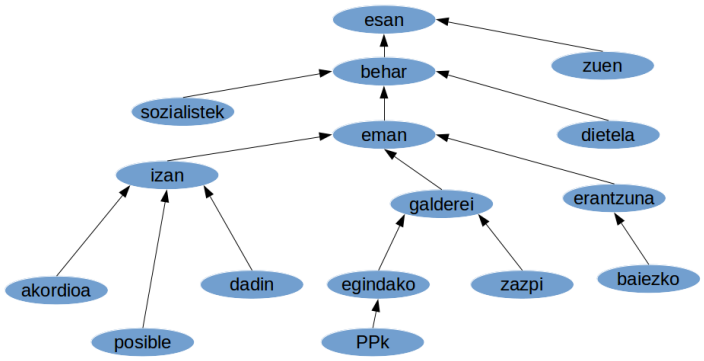
ii. Honako hau esan zuen: “Sozialistek PPk egindako zazpi galderei baiezko erantzuna eman behar diete. Akordioa posible izatea nahi da.”

(49c) multzoan, aldiz, dependentzia-zuhaitza goitik behera aztertu dugu (4.3 irudia). Aditz nagusitik gertuen dagoen fenomenoa sinplifikatu dugu

[5] Erreg.: perpausak banatu; erlazio-markak ezabatu; mendeko perpausaren aditza nominalizatu; *behar izan* txertatu; eta nagusia mendekoaren aurretik (nagusia$_{jat}$-mendekoa$_{jat}$) eman.

[6] Erreg.: zehar-estilotik estilo zuzenera pasa.

lehenik (perpaus konpletiboa, 4.4 irudia) eta (49c-i) esaldia lortu dugu. Ondoren dependentzia-zuhaitzean behera egin dugu eta helburu-perpausaren erregela aplikatu dugu (4.5 irudia). Emaitza (49c-ii) esaldia da.

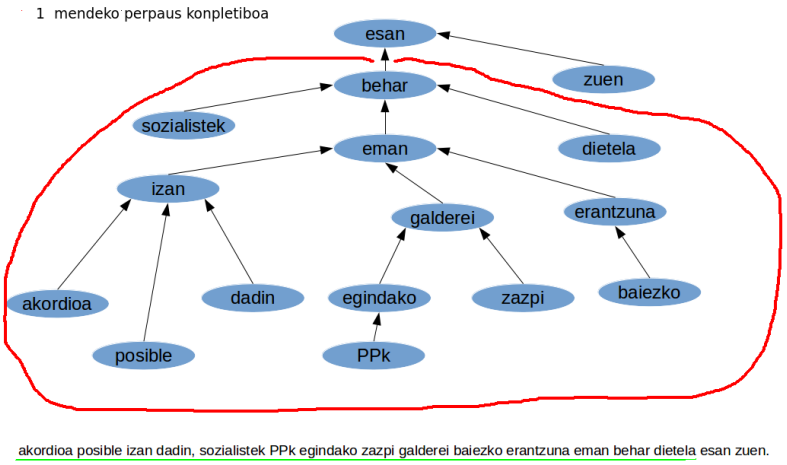


4.3 irudia – (49a) esaldiaren dependentzia-zuhaitza

Bi esaldi horiek (49b-ii) eta (49c-ii) erkatzen baditugu, badirudi ez direla oso ezberdinak. Baina, komatxoen artean geratzen den testua ezberdina da. (49b-ii) adibideko bigarren esaldia komatxoetik kanpo geratzen da, eta uler liteke esaldi hori ez duela esan duenak esan; (49c-i) eta (49c-ii) esaldietan, ordea, argi geratzen da esandakoa komatxoen barnean geratzen delako.

Dependentzia-zuhaitzaren sakoneraren arabera (*top-down*) erregelak nola aplikatu behar diren banan-banan, (50) adibidean ikusiko dugu. Gure aukera sendotzeko, (50g) multzoan fenomenoak agertu ahalako sinplifikazioaren emaitza eman dugu[7]. (50a) multzoan jatorrizko esaldia aurkeztu dugu eta (50f) multzora arteko adibide multzoetan aplikatutako erregelen emaitzak ikusiko ditugu. Hau da, lehendabiziko erregela aplikatu ondoren, (50b) multzoko esaldiak izango dira emaitza; bigarren erregelak aplikatuta, (50c)

[7]Esanahia asko aldatzen ez den arren, badira aldatzen diren ñabardurak. Kasu horretan agintariek azpimarratutakoak soilik bi esaldi dira eta ez esaldi guztiak.



4.4 irudia – (49a) esaldiaren dependentzia-zuhaitza, perpaus konpletiboa markatuta

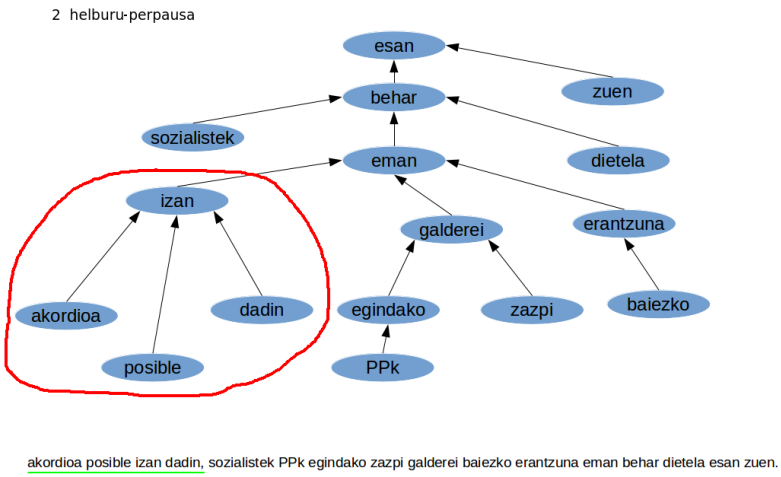
multzoa, eta horrela (50f) multzora iritsi arte. Unean lantzen ari garen fenomenoa beltzez adierazi dugu, argiago gera dadin.

Jatorrizko (50a) esaldiaren dependentzia-zuhaitzaren maila gorenean bi perpaus koordinatu ditugu. Horietako bigarrenari aditza falta zaionez, parentesi artean gehitu diogu. Orduan, aplikatu dugun lehendabiziko erregela[8] emendiozko perpaus koordinatuei dagokiena da, eta (50b) multzoko (50b-i) eta (50b-ii) esaldiak lortu ditugu. Perpaus koordinatuak nahiko independenteak direnez, koordinatu bakoitzak bere ibilbidea jarraituko du.

Perpaus koordinatuetan, aditz nagusitik gertuen dauden perpausak konpletiboak dira, eta horien erregela[9] aplikatu dugu (50c) multzoko adibideetan. (50c-ii) eta (50c-iv) adibideetan oraindik sinplifikatzeko fenomenoak daudenez, falta diren fenomeno horien erregelak aplikatu ditugu (50d) multzora igarotzeko.

(50c-ii) esaldian denbora-perpausaren[10] erregela aplikatu dugu, eta (50d)

[8]Erreg.: perpausak banatu eta juntagailua ezabatu.
[9]Erreg.: zehar-estilotik estilo zuzenera pasa.
[10]Erreg.: perpausak banatu, erlazio-markak ezabatu; *Orduan* TEa gehitu, eta mendeko



4.5 irudia – (49a) esaldiaren dependentzia-zuhaitza, helburu-perpausa markatuta

multzoko (50d-ii) eta (50d-iii) esaldiak lortu ditugu. (50c-iv) esaldian, berriz, perpaus erlatibo arrunten erregela[11] aplikatu dugu, (50d-v) eta (50d-vi) esaldiak lortzeko. Azken bi esaldi horietan jada ez dugu sinplifikatzeko fenomenorik, eta hemendik aurrera horrela mantenduko dira.

(50d-ii) esaldian, aldiz, badaude oraindik sinplifika daitezkeen fenomenoak. Goien dagoen fenomenoa kontzesio-perpausa da, eta horien erregela[12] aplikatu dugu (50e) multzoko (50e-ii) eta (50e-iii) esaldiak lortzeko. (50e-ii) esaldian, emendiozko perpaus koordinatu bat dagoenez, perpaus koordinatuen erregela[13] aplikatuta lortu ditugu (50f) multzoko (50f-ii) eta (50f-iii) esaldiak.

(50f) multzoko esaldietan jada sinplifikatzeko fenomeno gehiago ez dagoe-

perpausa nagusiaren aurretik (mendekoa$_{jat}$-nagusia$_{jat}$) eman.

[11]Erreg.: perpausak banatu, aurrekaria bi esaldi berrietan eman, eta mendekoa$_{jat}$-nagusia$_{jat}$ hurrenkera jarraitu.

[12]Erreg.: perpausak banatu, erlazio-markak ezabatu; *Hala ere* TEa gehitu, eta mendeko perpausa nagusiaren aurretik (mendekoa$_{jat}$-nagusia$_{jat}$) eman.

[13]Erreg.: perpausak banatu eta juntagailua ezabatu.

nez, bukatutzat eman dugu jatorrizko (50a) esaldiaren sinplifikazio sintaktikoa.

(50) a. Haize handia ibili eta euria egiten zuen arren, hegazkinari bidaia hasteko baimena eman zitzaionean eguraldia hegan egiteko modukoa zela azpimarratu zuten Taiwango Babes Zibileko agintariek, **eta** Boeing 747a bota zuen haize bolada bereziki indartsu hura ezin zela aurreikusi (azpimarratu zuten).

b. i. Haize handia ibili eta euria egiten zuen arren, hegazkinari bidaia hasteko baimena eman zitzaionean eguraldia hegan egiteko modukoa z**ela** azpimarratu zuten Taiwango Babes Zibileko agintariek.

ii. Boeing 747a bota zuen haize bolada bereziki indartsu hura ezin z**ela** aurreikusi (azpimarratu zuten).

c. i. Honako hau azpimarratu zuten Taiwango Babes Zibileko agintariek:

ii. "Haize handia ibili eta euria egiten zuen arren, hegazkinari bidaia hasteko baimena eman zitzaio**nean** eguraldia hegan egiteko modukoa zen."

iii. Honako hau ere azpimarratu zuten:

iv. "Boeing 747a bota zue**n** haize bolada bereziki indartsu hura ezin zen aurreikusi."

d. i. Honako hau azpimarratu zuten Taiwango Babes Zibileko agintariek:

ii. "Haize handia ibili eta euria egiten zue**n arren**, hegazkinari bidaia hasteko baimena eman zitzaion.

iii. Orduan eguraldia hegan egiteko modukoa zen."

iv. Honako hau ere azpimarratu zuten:

v. "Boeing 747a bota zuen haize boladak.

vi. Haize bolada bereziki indartsu hura ezin zen aurreikusi."

e. i. Honako hau azpimarratu zuten Taiwango Babes Zibileko agintariek:

ii. "Haize handia ibili **eta** euria egiten zuen.

iii. Hala ere, hegazkinari bidaia hasteko baimena eman zitzaion.

iv. Orduan eguraldia hegan egiteko modukoa zen."

v. Honako hau ere azpimarratu zuten:
vi. “Boeing 747a bota zuen haize boladak.
vii. Haize bolada bereziki indartsu hura ezin zen aurreikusi.”

f. i. Honako hau azpimarratu zuten Taiwango Babes Zibileko agintariek:
ii. “Haize handia ibili zen.
iii. Euria egiten zuen.
iv. Hala ere, hegazkinari bidaia hasteko baimena eman zitzaion.
v. Orduan eguraldia hegan egiteko modukoa zen.”
vi. Honako hau ere azpimarratu zuten:
vii. “Boeing 747a bota zuen haize boladak.
viii. Haize bolada bereziki indartsu hura ezin zen aurreikusi.”

g. i. Haize handia ibili zen.
ii. Euria egiten zuen.
iii. Hala ere, hegazkinari bidaia hasteko baimena eman zitzaion.
iv. Orduan honako hau azpimarratu zuten Taiwango Babes Zibileko agintariek:
v. “Eguraldia hegan egiteko modukoa zen.”
vi. Boeing 747a bota zuen haize bolada haize boladak.
vii. Honako hau ere azpimarratu zuten:
viii. “Haize bolada bereziki indartsu hura ezin zen aurreikusi.”

(50) adibidean dependentzia-zuhaitza goitik behera zeharkatu dugu. (50g) multzoan esaldiaren sinplifikazioaren emaitza eman dugu ezkerretik eskuinerako hurrenkerari jarraituz, bi emaitzen arteko ezberdintasunak ikus daitezen.

Baina zer gertatzen da maila berean egitura batekin baino gehiagorekin topo egiten badugu? Egitura horietako fenomenoak ezberdinak badira, aldi berean sinplifikatuko ditugu. (51) adibidean maila berean perpaus erlatiboa eta aposizio-sintagma ditugu, eta horiek erregela-multzo ezberdinetakoak[14] direnez ez dugu arazorik batera sinplifikatzeko. Perpaus erlatiboei dagokien

[14]Erregela-multzoak fenomenoen arabera bereizten dira: koordinazioko erregelak, perpaus erlatiboen erregelak, perpaus adberbialen erregelak, perpaus osagarrien erregelak, aposizioen erregelak eta egitura parentetikoen erregelak. Multzo horiek B eranskineko tauletan ikus daitezke.

erregela[15] (51b-i) eta (51b-ii) esaldiak sortzeko erabili dugu, eta aposizioen erregela[16] (51b-ii), (51b-iii) eta (51b-iv) esaldiak lortzeko aplikatu dugu.

(51) a. IPAR IRLANDAKO bake prozesuak ditu**en** bi arazoren inguruan jarrera bateratua hartzeko asmoz bildu ziren atzo **Londresen Tony Blair Erresuma Batuko lehen ministroa** eta **Bertie Ahern Irlandako Errepublikakoa**.

b. i. Ipar Irlandako bake prozesuak bi arazo ditu.

ii. Bi arazo horien inguruan jarrera bateratua hartzeko asmoz bildu ziren atzo Londresen Tony Blair eta Bertie Ahern.

iii. Tony Blair Erresuma Batuko lehen ministroa da.

iv. Bertie Ahern Irlandako Errepublikakoa da.

Beraz, esaldi batean erregelak aplikatzeko dependentzia-zuhaitzean goien dauden egiturekin (52) hasi gara eta behera jarraitu dugu. Maila berean bi fenomeno ezberdin aurkitu ondoren, horiek batera sinplifikatu ditugu eta dependentzia-zuhaitzean behera jarraitu dugu. Lehenik, i) perpaus koordinatuaren erregela[17] aplikatu dugu (B multzoa). Ondoren, ii) koordinatu batean maila berean perpaus konpletiboa eta denbora-perpausa ditugunez, bien erregelak[18] aplikatu ditugu batera (C multzoa). Esaldi horietan behera jarraitu dugu, eta iii), horietako batean aposizioa aurkitu dugunez, aposizioaren erregela[19] aplikatu dugu (D multzoa).

(52) a. 30 urtekoa, Indonesiako ringetan zaildua dago Victor Ramos, **eta** dominen borrokan sartzeko helburuz zetorre**la** adierazi zuen **Eki Timorreko ordezkaritzaburu Frank Fowliek**, taldea Sydneyra heldu ze**nean**.

[15]Erreg.: perpausak banatu, aurrekaria bi esaldi berrietan eman, eta mendekoa$_{jat}$-nagusia$_{jat}$ hurrenkera jarraitu.

[16]Erreg.: aposizioan dauden sintagmak banatu, kargua adierazten duen sintagma perpaus nagusitik ezabatu, aposizioan dauden sintagmekin esaldi berriak sortu eta esaldi berriak nagusia$_{jat}$-mendekoa$_{jat}$ (mendekoak, kasu honetan, aposizio-sintagmetatik sortutako esaldiak dira) hurrenkera jarraituz eman.

[17]Erreg.: perpausak banatu eta juntagailua ezabatu.

[18]Perpaus konpletiboaren erreg.: zehar-estilotik estilo zuzenera pasa.
Denbora-perpausaren erreg.: perpausak banatu, erlazio-markak ezabatu; *Orduan* TEa gehitu eta mendeko perpausa nagusiaren aurretik mendekoa$_{jat}$-nagusia$_{jat}$ eman.

[19]Erreg.: aposizioan dauden osagaiak (bi izen-sintagmak) banatu; hondarkia esalditik ezabatu; bi osagaiekin esaldi kopulatiboa sortu.

b. i. 30 urtekoa, Indonesiako ringetan zaildua dago Victor Ramos.
   ii. Dominen borrokan sartzeko helburuz zetorrela adierazi zuen Eki Timorreko ordezkaritzaburu Frank Fowliek, taldea Sydneyra heldu zenean.

c. i. 30 urtekoa, Indonesiako ringetan zaildua dago Victor Ramos.
   ii. Taldea Sydneyra heldu zen.
   iii. Orduan, honako hau adierazi zuen:
   iv. "Dominen borrokan sartzeko helburuz dator Eki Timorreko ordezkaritzaburu Frank Fowlie."

d. i. 30 urtekoa, Indonesiako ringetan zaildua dago Victor Ramos.
   ii. Taldea Sydneyra heldu zen.
   iii. Orduan, honako hau adierazi zuen:
   iv. "Dominen borrokan sartzeko helburuz dator Frank Fowlie."
   v. Frank Fowlie Eki Timorreko ordezkaritzaburua da.

Hierarkia horretan, hala ere, bi murriztapen ezarri behar izan ditugu:

1. Esaldi batean ezin da bi aldiz perpaus osagarrien erregela aplikatu, betiere koordinatu ezberdinak ez badira. Hau da, perpaus batean perpaus osagarrien erregela aplikatu badugu, horren mende beste perpaus osagarri bat topatzen badugu, azken hori ez dugu aplikatuko, sinplifikatuko alegia. Horrekin, *honako hau* txertatze-elementua gehiegi erabiltzeaz gain, estilo zuzen barruko estilo zuzenak ekidin nahi ditugu. Hau da, ez dugu nahi *Honako hau uste dut: "Honako hau iruditzen zitzaidan: "Oso azkar nindoan.""* bezalako esaldirik sortu. Arazo hori (53) adibidean erakusten dugu.

(53) a. Beldur nintzen azken metroetan hanketan indarrik ez ote nuen izango, **eta**, oso azkar ez nindo**ala** iruditzen zitzaida**n arren**, nahikoa ondo joan d**ela** uste dut.

b. i. Beldur nintzen azken metroetan hanketan indarrik ez ote nuen izango.
   ii. Honako hau uste dut:
   iii. "Oso azkar ez nindoala iruditzen zitzaidan.
   iv. Hala ere, nahikoa ondo joan da."

2. Esaldi berean ere (edo perpaus koordinatuak banatu ondoren sortutako perpausetan), perpaus adberbialen erregelak ez ditugu bitan baino gehiagotan aplikatuko. Horrekin txertatze-elementuen gehiegizko erabilera ekidin nahi dugu, horiek testuaren erritmoa hauts dezaketelako.

Zein erregela aplikatuko ditugun jakiteko, dependentzia-zuhaitzaren hierarkia aplikatuko dugu, eta goien daudenak izango dira sinplifikatuko direnak. (54) adibideko (54a) jatorrizko esaldian mendeko hiru perpaus adberbial daude. Goien daudenak baldintza-perpausa eta kausa-perpausa direnez, horien erregelak[20] aplikatu ditugu.

(54) a. Erregularitatearen saririk eman **ba**lute, hura bere-berea izango zen seguru asko, hasiera beretik hartu **bait**zuen altura, esperientzia handiagoko bertsolariak kili-kolo abiatu zire**n bitartean**.

b. i. Erregularitatearen saririk ez dute ematen.

ii. Bestela, hura bere-berea izango zen seguru asko.

iii. Izan ere, hasiera beretik hartu zuen altura, esperientzia handiagoko bertsolariak kili-kolo abiatu ziren bitartean.

Erregelak zein hurrenkeratan eman behar diren ere TSAko beste lanetan aztertu dute. Ingelesaren kasuan, erregelen aplikazioa eskuz definitu dute transformazio ohikoetan oinarrituz (Chandrasekar *et al.*, 1996); ezkerretik eskuinerako hurrenkera ere jarraitu dute, esperimentuetan emaitza onenak eman dituztelako (Siddharthan, 2002) edo, gurean bezala, goitik beherako aplikazioa egin dute (Siddharthan, 2006). Brasilgo portugesaren kasuan, berriz, teilakatuta aplikatzen dituzte erregelak hurrenkera honi jarrikiz: ahots pasiboa, aposizioa, mendeko perpausak, perpaus erlatibo ezmurriztailea, perpaus erlatibo murriztailea eta perpaus koordinatua (Gasperin *et al.*, 2009a). Erdal hizkuntzetako hurrenkera horiek 4.2 taulan laburtu ditugu.

Orain arte aurkeztu ditugun ezaugarriak egun dauzkagun tresnekin inplementatzeko gai gara. Badira, hala ere, oraindik inplementatzeko gai ez garen

[20] Baldintza-perpausaren erreg.: perpausak banatu, erlazio-markak ezabatu; mendeko perpausean *Bestela* TEa gehitu eta polaritatea aldatu; mendekoa$_{jat}$-nagusia$_{jat}$ hurrenkera jarraitu.
Kausa-perpausaren erreg.: perpausak banatu, erlazio-markak ezabatu; *Izan ere* TEa gehitu eta nagusia$_{jat}$-mendekoa$_{jat}$ hurrenkera jarraitu.

Hizkuntza	Lana	Erregelen hurrenkera
Ingelesa	(Chandrasekar *et al.*, 1996)	Eskuz definituta
	(Siddharthan, 2002)	Ezkerretik eskuinera
	(Siddharthan, 2006)	Goitik behera
Brasilgo portugesa	(Gasperin *et al.*, 2009a)	Teilakatuta

4.2 taula – Sinplifikazio-erregelen hurrenkera beste hizkuntzetan

ezaugarri batzuk: korreferentziaren ebazpena, elipsiaren tratamendua eta *Wikipedia*rako estekak. Horiek etorkizunean integratzekoak izango dira.

Korreferentziaren ebazpenarekin, testuan agertzen diren izenordainak dagokien sintagmarekin ordezkatuko ditugu. Elipsiaren tratamenduan eliditutako argumentuak berreskuratuko ditugu eta *Wikipedia*rako estekak eginez entitateei edo gertaerei buruzko informazio gehigarria emango dugu. Hiru ezaugarri horiekin testua aberastuko dugu, beste mota batzuetako sinplifikazioetara jauzi eginez, semantikoa edo azalpenezkoa, esaterako. Halaber, kontuan izan behar dugu testuen sinplifikazio automatikoan sinplifikazio lexikala ere landu dela (ikus 2. kapitulua).

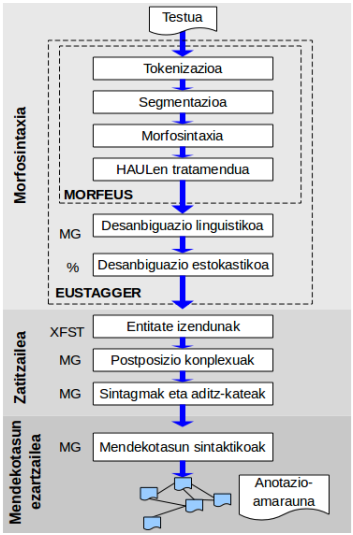
4.3 Analisi automatikorako tresnak

Testuen konplexutasuna aztertzeko eta testu konplexu horien sinplifikazio sintaktiko automatikoa gauzatzeko ezinbestekoa da testua automatikoki analizatzea. Atal honetan euskarazko testuak automatikoki analizatzeko behar ditugun tresnak azalduko ditugu. Tresna horietako batzuk Ixa taldearen analisi-katean integratuta daude eta Mugak eta aposizio-detektatzailea sinplifikazio automatikoaren beharrez moldatu edo sortu ditugun tresnak dira; azken horiek oraindik analisi-katean integratzeke daude. Tresna horiek sinplifikazio sintaktikoko erregelak inplementatzean zertarako izango diren erabilgarriak ere aipatuko dugu.

4.3.1 Ixa taldearen analisi-katea

Ixa taldearen analisi-katea Ixa taldean testuak analizatzeko erabiltzen diren oinarrizko tresnen multzoa da (4.6 irudia). Tresna horien oinarri lexikala EDBL-LBDBL (Euskararen Datu-Base Lexikala - Lexikoaren Behatokiko Datu-Base Lexikala) da. EDBL-LBDBL baliabide lexikal konputazionala da (Aldezabal *et al.*, 2001). Hasiera batean Xuxen zuzentzaile ortografikoaren

euskarri bezala garatu bazen ere, egun Morfeus analizatzaile morfologikoaren eta Euslem lematizatzailearen oinarri lexikala da. EDBL-LBDBLren informazioa hiru atal nagusitan banatzen da: i) hiztegi-sarrerak (hiztegi konbentzional batean aurkitzen diren bezala), ii) aditz-formak eta bestelako forma flexionatuak eta iii) morfema ez-independenteak. Sarrera bakoitzak bere informazio morfologikoa du. Azken urteetan, Euskaltzaindiaren Lexikoaren Behatokiko eta Hiztegi Batuko hitzekin aberastu da. Egun[21], 122.383 sarrera ditu (horietatik 101.423 hiztegi-sarrera), eta Hiztegi Batuan arautuak dauden ere jasotzen du.



4.6 irudia – Ixa taldearen analisi-katea

EDBL-LBDBLn oinarrituz, Ixa taldearen analisi-katea (4.6 irudia) osatzen duten tresnak honako hauek dira:

- Morfeus analizatzaile morfologikoa (Aduriz *et al.*, 1998) (4.7 irudia): Tresna honek ematen dio hasiera analisi linguistikoari. Morfeusek sarrerako testua jasotzen du eta tokenetan banatzen du. Ondoren, token horiei lema eta konbinazio posible guztiak, eta informazio morfologikoa esleitzen dizkie. Zehazki, ondorengo prozesu hauek gauzatzen ditu:

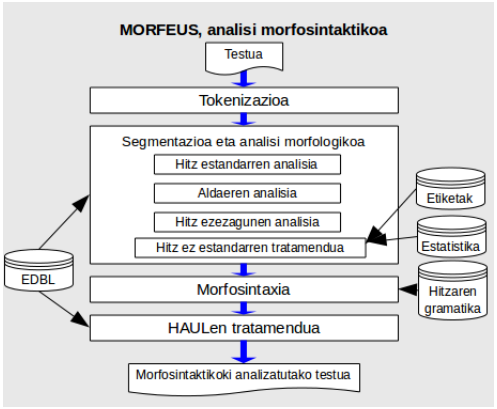
[21] Datuak 2016ko urtarrilaren 12an erauzi ditugu.

– Tokenizazioa: Testua tokenetan eta esaldietan banatzen du. Gainera, testuan hitzak, zenbakiak (arruntak zein erromatarrak; deklinatu gabeak edo deklinatuak), laburdurak, siglak, zuriuneak eta puntuazio-markak identifikatzen ditu. Tokenei ere informazio geografikoa (identifikadorea) gehitzen die.

– Segmentazio edo analisi morfologikoa (Alegria, 1995; Ezeiza, 2002; Urkia, 1997): Segmentatzaileak hitz bakoitza lemetan eta morfemetan banatzen du eta osagai horien morfotaktika (morfemen arteko konbinazio posibleak) eta informazio morfologikoa ematen du. EDBLko informazioa erabiltzen du eta hitz ez estandarren tratamenduan etiketak eta estatistika erabiltzen ditu.

– Morfosintaxia (Aduriz *et al.*, 2000; Gojenola, 2000): Segmentatzaileak emandako informazioa bildu eta optimizatu egiten du informazioa garbituz, eta hitz-formari bere osotasunean dagokion informazio morfologikoa esleitzen dio. Horretarako hitzaren gramatika erabiltzen du.

– Hitz Anitzeko Unitate Lexikalen (HAUL) tratamendua (Alegria *et al.*, 2004; Urizar, 2012): Hitz elkartuen, lokuzioen eta kolokazio murriztuen tratamendua egiten du. Horretarako, EDBLko informazioa erabiltzen du.



4.7 irudia – Morfeus analizatzaile morfologikoa

Morfeusen irteera[22] (45) adibideko lehenengo perpausarekin ikus

[22]Irudietan agertzen diren etiketen adierak Ixa Taldearen etiketen eskuliburuan (Ixa

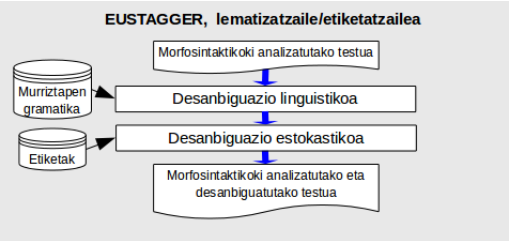
dezakegu, 4.8 irudian. Morfeusek hitz bakoitzari ezaugarri morfologiko guztiak esleitzen dizkio.

4.8 irudia – (45) adibideko esaldia, Morfeus tresnaren irteerarekin

Morfeus izango da erregeletan morfemen banaketa egiteko erabiliko dugun tresna.

- Eustagger lematizatzaile/etiketatzailea (Ezeiza, 2002; Aduriz, 2000; Aduriz Taldea, 2004) aurki daitezke.

riz eta Díaz de Ilarraza, 2003; Aduriz *et al.*, 1997) (4.9 irudia): Morfeusen irteera sarrera izanda, hitz bakoitzari, dagokion testuinguruan, lema eta etiketak esleitzen dizkio. Anbiguotasuna ebazteko, ezagutza linguistikoa (murriztapen-gramatika) eta ezagutza estatistikoa (etiketen bidezko desanbiguazio estokastikoa) erabiltzen ditu.



4.9 irudia – Eustagger lematizatzaile/etiketatzailea

Eustaggerren irteera (45) adibideko esaldiarekin ikus dezakegu 4.10 irudian. Eustaggerrek hitz bakoitza ezaugarri morfologiko bakarrarekin uzten du behin desanbiguatuta.

Asperren	**kasua**	**emeki-emeki**	**aitzinatu**	**bada**	**ere**	**,**
Asper	*kasu*	*emeki-emek*	*aitzinatu*	*izan*	*ere*	*,*
IZEIZB	IZEARR	IZEARR	ADI	ERLMEN	LOTLOK	PUNT_KOMA

Sa	**Pintoren**	**etorkizuna**	**fite**	**argituko**	**da**	**.**
sa	*pinto*	*etorkizun*	*fite*	*argitu*	*izan*	*.*
IZEARR	ADJIZO	IZEARR	ADB	ADI	ADL	PUNT_PUNT

4.10 irudia – (45) adibideko esaldia Eustagger tresnaren irteerarekin

Eustaggerren irteerak adieraziko digu zein motatakoa den analizatzen ari garen perpausa. Horretaz gain, aditzaren informazio morfologikoa desanbiguatuta emango digu, aditza jokatua edo ez-jokatua den, aldia, aspektua eta pertsonak.

- Ixati zatitzailea edo *chunker*ra (Aduriz *et al.*, 2004) (4.11 irudia): Testua kateetan (*chunk*etan[23]) banatzen diren erlazionaturiko hitz-multzoak hauek dira:

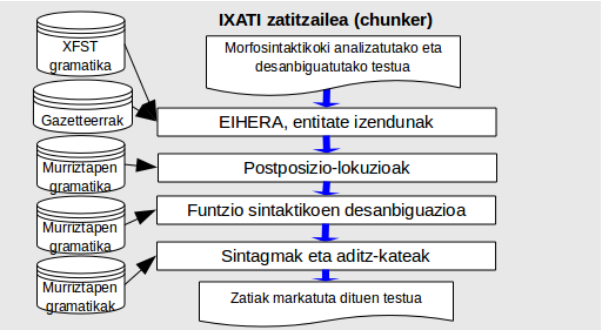
[23] *Chunk*ak buru bat duten eta gainjartzen ez diren osagaien continuuma dira (Abney, 1991).

– Entitate izendunak (EIHERA) (Alegria *et al.*, 2003; Fernandez Gonzalez, 2012): Entitate-izenak (pertsonak, tokiak eta erakundeak), denbora-adierazpenak (datak eta orduak), eta zenbait zenbakizko espresio (portzentajeak, diru-balioak...) identifikatu eta sailkatzen ditu. Horretarako XSFT gramatikak eta gazettererrak erabiltzen ditu.

– Postposizio- eta menderagailu-lokuzioak: Murriztapen-gramatika baten bidez ezagutzen ditu (Urizar, 2012).

– Funtzio sintaktikoen (FS) desanbiguazioa (Aduriz, 2000): Analisi batek funtzio sintaktiko bat baino gehiago badu, murriztapen-gramatiken bitartez desanbiguatzen ditu esaldiaren azaleko sintaxiaren berri emateko.

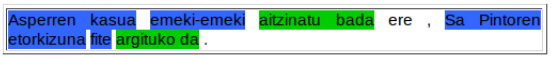
– Sintagmak eta aditz-kateak (Aranzabe, 2008; Arrieta, 2010): Sintagmen barruan beste sintagmarik edo katerik duten hitz-multzoak aurkitzen ditu (ezagutza linguistikoan oinarritutako murriztapen-gramatika eta ezagutza estatistikoan oinarritutako tresna).



4.11 irudia – Ixati zatitzailea

Ixatiren irteera (45) adibideko esaldiarekin 4.12 irudian ikus dezakegu. Koloreek zatiak zatiak adierazten dituzte: urdinak izen-sintagmak eta berdeak aditz-sintagmak.

Ixatiren irteeratik batez ere entitateak, postposizio-lokuzioak, eta sintagma eta aditz-kateak erabiliko ditugu. Entitateak aposizioen erregeletan erabiliko ditugu gehienbat, postposizio-lokuzioen identifikazioa landu ditugun postposizio-egiturak non banatu behar diren jakiteko,



4.12 irudia – (45) adibideko esaldia Ixati tresnaren irteerarekin

eta sintagmen eta aditz-kateen etiketak luzera minimoa egiaztatzeko baliatuko ditugu.

- Dependentzia-erlazioen ezartzailea: Hitzak binaka lotzen ditu esaldiaren zuhaitz sintaktikoa (dependentzia-zuhaitza) lortzeko. Euskaraz hiru tresna daude mendekotasunak ezartzeko.
  - EDGK (Euskarako Dependentzia Gramatika Konputazionala) (Aranzabe, 2008): Ezagutza linguistikoan oinarrituta, bi urratsetan ezartzen ditu mendekotasunak. Lehenik, mendeko elementuari gobernatzailearekiko duen erlazioa ezartzen dio eta, bigarrenik, gobernatzailea zein token den adierazten du.
  - *Maltparser/Maltixa* (Bengoetxea eta Gojenola, 2007; Bengoetxea, 2014): estatistikan oinarrituta dependentzia-erlazioak ezartzen ditu.
  - Analizatzaile hibridoa (Aranzabe *et al.*, 2012b): ezagutza linguistikoaren eta estatistikoaren konbinazioan oinarrituta dependentzia-erlazioak ezartzen ditu.

    Maltixaren irteera (45) adibideko esaldiarekin 4.13 irudian ikus dezakegu. Hirugarren zutabeko zenbakiak adierazten du lerro horretako hitzak esaldiko zenbatgarren hitza duen gobernatzailea. Laugarren zutabean adierazten da bi hitzen arteko erlazio-mota zein den.

  Dependentzia-erlazioen ezartzailea erregelen hurrenkera aplikatzeko erabiliko dugu.

Analisi-kate horretan testu bat automatikoki aztertzeko erabiltzen diren tresnez gain, testuen konplexutasunaren analisirako eta testuen sinplifikazio automatikorako beste bi dira beharrezkoak:

- Mugak (Ondarra, 2003; Aduriz *et al.*, 2006b; Arrieta, 2010): Esaldien eta perpausen mugak ezartzen dituzten bi tresna daude euskaraz:

Syntax analysis of: Asperren kasua emeki-emeki aitzinatu bada ere, Sa Pintoren etorkizuna fite argituko da. is:

Index	Word	Head	Relation	Lemma	Features	Category	Subcategory
1	Asperren	2	ncmod	asper	KAS:GEN|NUM:P	IZE	IZE_ARR
2	kasua	4	ncsubj	kasu	KAS:ABS|NUM:S	IZE	IZE_ARR
3	emeki-emeki	4	ncmod	emeki-emeki	_	ADB	ADB_ARR
4	aitzinatu	0	ROOT	aitzinatu	ADM:PART|ASP:BURU	ADI	ADI_SIN
5	bada	4	auxmod	izan	ERL:BALD	ADL	ADL
6	ere	5	mw	ere	_	LOT	LOT_LOK
7	,	6	PUNC	,	_	PUNT	PUNT_KOMA
8	Sa	10	ncmod	Sa	_	IZE	IZE_IZB
9	Pintoren	8	ncmod	pinto	KAS:GEN	ADJ	ADJ_ARR
10	etorkizuna	12	ncsubj	etorkizun	KAS:ABS|NUM:S	IZE	IZE_ARR
11	fite	12	ncmod	fite	_	ADB	ADB_ARR
12	argituko	0	ROOT	argitu	ADM:PART|ASP:GERO	ADI	ADI_SIN
13	da	12	auxmod	izan	_	ADL	ADL
14	.	4	PUNC	.	_	PUNT	PUNT_BEREIZ

4.13 irudia – (45) adibideko esaldia Maltixa tresnaren irteerarekin

  - Ezagutza linguistikoan oinarritutakoa: Murriztapen-gramatikaren eredua jarraituz gramatika batek esaldien mugak jartzen ditu, eta beste gramatika batek perpausen mugak.
  - Estatistikan oinarritutakoa: ezagutza linguistikoan oinarritutako gramatikak oinarri hartuta, teknika estatistikoak aplikatuz esaldien eta perpausen mugak ezartzen ditu.

    Mugak perpausen mugak identifikatzeko erabiliko ditugu; beraz, perpausak non banatu behar diren adieraziko digu.

- Aposizio-detektatzailea: Murriztapen-gramatikak erabiliz, aposizio hautagaiak markatzen ditu eta ondoren aposizio-sintagmak berresten ditu. Aposizio-detektatzailearen bidez aposizio-sintagmen mugak eta aposizio-motak zein diren ezagutuko dugu. Horiek erabiliz, aposizio-sintagmak non banatu behar diren jakingo dugu.

Prozesu horien guztien irteera 4.14 irudian ikus daiteke. (45) esaldia adibide hartuta, lemak marroiz azpimarratu ditugu; analisi morfologikoa, gorriz; entitateak, lilaz; funtzio sintaktikoak, berde argiz; azaleko sintaxia edo zatiak, urdinez; esaldi-mugak, horiz; perpaus-mugak, urdin argiz; eta dependentziak fuksiaz. Adibide horretan ez dugu aposiziorik topatu.

Konplexutasunaren azterketarako eta testuen sinplifikazio automatikoa gauzatzeko analisi-prozesu horretan hobetu eta garatu ditugun Mugak eta Aposizioak tresnak hurrengo azpiataletan deskribatuko ditugu.

```
"<Asperren>"<HAS_MAI>"   S:137/0
        "asper" IZE ARR GEN NUMP MUGM ZERO HAS_MAI w1,L-A-IZE-ARR-10,lsfi2 @IZLG> %SIH S:137 %ESALDI_HAS_1_BEREZIA &NCMOD>
"<kasua>"
        "kasu" IZE ARR BIZ- ABS NUMS MUGM w2,L-A-IZE-ARR-14,lsfi3 @SUBJ %SIB &NCSUBJ>
"<emeki-emeki>"
        EZEZAG "emeki-emeki" ADB ARR ZERO w3,L-G-ADB-ARR-2,lsfi6 @ADLG %SINT &NCMOD>
"<aitzinatu>"
        "aitzinatu" ADI SIN PART BURU NOTDEK w4,L-A-ADI-SIN-8,lsfi7 @-JADNAG %ADIKATHAS &MENOS>
"<bada>"
        "izan" ADL BALD A1 NOR NR_HURA w5,L-A-ADL-3,lsfi8 @+JADLAG_MP_ADLG %ADIKATBU &MENOS>
"<ere>"
        "ere" LOT LOK EMEN w6,L-A-LOT-LOK-6,lsfi11 @LOK &CMOD>
"<,>"<PUNT_KOMA>"  S:608/0
       PUNT_KOMA S:608 }MUGA
"<Sa>"<HAS_MAI>"
        EZEZAG "Sa" IZE IZB PLU- ENTI_HAS_PER AORG HAS_MAI w8,L-G-IZE-IZB-3,lsfi12 @KM> %SIH &NCMOD>
"<Pintoren>"<HAS_MAI>"
        "pinto" ADJ ARR IZAUR- GEN MG ENTI_BUK_PER HAS_MAI w9,L-A-ADJ-ARR-10,lsfi14 @IZLG> &NCMOD>
"<etorkizuna>"
        "etorkizun" IZE ARR BIZ- ABS NUMS MUGM w10,L-A-IZE-ARR-18,lsfi15 @SUBJ> %SIB &NCSUBJ>
"<fite>"
        "fite" ADB ARR ZERO w11,L-A-ADB-ARR-3,lsfi18 @ADLG %SINT &NCMOD>
"<argituko>"
        "argitu" ADI SIN PART GERO NOTDEK w12,L-A-ADI-SIN-10,lsfi19 @-JADNAG %ADIKATHAS
"<da>"
        "izan" ADL A1 NOR NR_HURA w13,L-A-ADL-5,lsfi20 @+JADLAG %ADIKATBU &<AUXMOD
"<$.>"<PUNT_PUNT>"  S:123/0  S:148/0
       PUNT_PUNT S:123 %ESALDI_BUK_1 S:148 }MUGA
```

4.14 irudia – (45) adibideko esaldiaren analisi automatikoa

4.3.2 Mugak: MuGa gramatikaren egokitzapena

MuGa gramatikak perpausen amaierako mugak etiketatzen ditu eta, erregelak aplikatzeko, murriztapen-gramatika (MG) *Constraint Grammar* (CG) (Karlsson *et al.*, 1995) formalismoa erabiltzen du. Gure lana gramatika horren (Ondarra, 2003; Aduriz *et al.*, 2006b) bertsio-berritzea (analisi-kateko bertsio berrira) eta erregelen hobekuntza izan da. Aldaketak, batez ere, aditz konposatuek osatzen dituzten perpausekin eta komarekin egin ditugu. Une honetan 78 erregela ditu gramatikak.

Gramatika honek }MUGA etiketa perpausaren amaieran jartzen du. Adibide gisa, 4.15 irudian, denbora-perpausen mugak detektatzeko erregela bat ikus dezakegu. Erregela horretan adierazten dugu:

- Esleitu }MUGA etiketa (MAP agindua),
- analisian denborazkoa bada (TARGET DENB),
- eta ondoko baldintzak (IF) betetzen baditu:
  - uneko hitza (0), aditz trinkoa (ADT), aditz laguntzailea (ADL), aditz konposatua (ADK) edo aditz-izena (ADIZE) izan behar da,
  - eta hurrengo hitza (1), ez (NOT) da PUNTUAZIOA zerrendan (puntuazio-markak biltzen diren zerrenda) agertu behar.

```
MAP (}MUGA) TARGET (DENB) IF (0 ADT OR ADL OR ADK OR ADIZE)
                             (NOT 1 PUNTUAZIOA);
```

4.15 irudia – Denbora-perpausak detektatzeko MG erregela bat

Erregela horrek (55) adibidea bezalako esaldietan denbora-perpausen mugak detektatzen ditu.

(55) Horregatik, Guiñazuk Gasteizera etorri zenean`}MUGA` ez zuen eman beharreko guztia ematen.

Gramatika honen hobekuntza etengabeko lana da, egitura berriak aurkitu ahala eta corpus berriak probatzearekin batera osatzen doana. Etorkizunerako duen erronkarik handiena hurrenkera kanonikoa ez duten esaldien eta perpausen doitasuna areagotzean datza.

Gramatika egokitu ondoren, denbora-perpausekin eta erlatibozko perpausekin ebaluatu dugu (bien artean 22 erregela). MuGa gramatika ebaluatzeko EPEC corpusa erabili dugu. Corpusa garapen- eta ebaluazio-zatitan banatu dugu. Lehenengo zatia erregelak hobetzeko eta diseinatzeko erabili dugu eta bigarrena, aldiz, ebaluaziorako. Azken zati hori eskuz etiketatu dugu, eta urre-patroia sortu dugu. 4.3 taulan corpusaren zati bakoitzak duen hitz-, esaldi- eta perpaus-kopurua ikus daitezke.

	Garapen-zatia	**Ebaluazio-zatia**
Hitz-kopurua	61.121	63.766
Esaldi-kopurua	5.068	5.211
Perpaus-kopurua	18.301	18.356

4.3 taula – Hitz-, esaldi- eta perpaus-kopurua EPEC corpuseko laginetan

Erlatibozko perpausekin eta denbora-perpaus adberbialekin lortutako emaitzak 4.4 taulan ikus daitezke. Ebaluaziorako erabili diren neurriak doitasuna[24], estaldura[25] eta F-neurria[26] izan dira. Taularen 4. zutabean mota bakoitzaren perpaus-kopurua adierazi dugu. Elementu askeak, berriz, denbora-perpaus adberbialetan menderagailuen ondoren aske (hitz soltea) doazen adberbio eta izenak dira.

[24] Doitasuna = zuzen detektatutako perpausak / detektatutako perpausak
[25] Estaldura = zuzen detektatutako perpausak / perpaus guztiak
[26] F-neurria = 2 * doitasuna * estaldura / (doitasuna + estaldura)

	Doitasuna	**Estaldura**	***F*-neurria**	**Kopurua**
Erlatibozko perpaus jokatuak	0,998	0,978	0,988	547
Erlatibozko perpaus ez-jokatuak	1	0,985	0,992	335
Denbora-perpaus jokatuak	0,955	0,964	0,960	111
Denbora-perpaus ez-jokatuak	0,966	0,966	0,966	29
Denbora-perpaus jokatuak + elementu askeak	1	0,556	0,714	18
Denbora-perpaus ez-jokatuak + elementu askeak	0,970	0,372	0,538	86

4.4 taula – Ebaluatutako fenomenoen emaitzak eta perpaus-kopurua

Emaitzen azterketari ekingo diogu orain. Erlatibozko perpausen emaitzak altuak dira. Aditz jokatua dutenen *F*-neurria 0,988 da eta ez-jokatuena 0,992. Erroreak analizatu ditugu eta ikusi dugu: a) erregela batek elementu askeak dituen denbora-perpaus jokatuen erregela batekin talka egiten duela; b) analisi morfologikoan erroreak daudela; eta c) aditz modal ez-jokatuak ez zeudela garapen-zatian eta, ondorioz, ez zegoen erregelarik.

Denbora-perpausen emaitzak bi taldetan banatu ditugu: elementu askeak ez dituztenak eta elementu askeak dituztenak. Lehenengo taldeko emaitzak oso antzekoak dira perpaus jokatuetan (*F*-neurria: 0,960) eta ez-jokatuetan (*F*-neurria: 0,966). Errore-analisia egin dugu, eta gehienak hurrenkera kanonikoa ez den beste bat erabiltzeagatik gertatzen dira. Bigarren taldeko emaitzak, ordea, baxuagoak dira. Denbora-perpaus jokatuen *F*-neurria 0,714 da eta ez-jokatuena 0,538. Talde honen arazoa da estaldura oso baxua dela (jokatuak 0,556 eta ez-jokatuak 0,372) eta horren arrazoia da, batetik, elementu askeak anbiguoak direla; eta bestetik, egitura horiek oso aberatsak direla eta ez ditugula denak garapen-zatian aurkitu. Hala ere, anbiguetateaz gain, doitasuna oso altua da (jokatuak 1 eta ez-jokatuak 0,970).

Gure helburua doitasuna lortzea denez (ez baititugu esaldiak okerreko puntuan banatu nahi) emaitza horiekin aise lortu dela uste dugu. Ondorioz, pentsatzen dugu MuGa gramatika testuen aurreprozesua egiteko tresna egokia dela, eta sinplifikazio sintaktikoaren prozesuaren oinarri egokia dela, hau da, perpausak non banatu behar diren jakiteko baliagarria zaigula.

4.3.3 Aposizioak: aposizio-detektatzailea

Aposizio-sintagmek izen-sintagma luzeagoak egiten dituzte, eta beste hizkuntzetako lanetan eta gure azterketa linguistikoan horiek sinplifikatzeko beharra ikusi dugu. Ikasketa automatikoarekin egindako esperimentuetan ere ikusi dugu ezaugarri esanguratsuenen artean daudela (ikus 5. kapitulua). Ixa taldearen analisi-katean ez zegoen aposizio-sintagmak eta horien mugak detektatzen zituen tresnarik; hori dela-eta, aposizio-detektatzailea garatu dugu. Aposizio-detektatzailea MG formalismoa erabiltzen duten bi gramatiketan oinarritzen da.

Aposizio-sintagmak detektatzeko bi fase ditugu, eta gramatika bakoitzak bat gauzatzen du. Lehenengo gramatikak aposizio-sintagma izateko hautagai diren entitateak markatzen ditu, eta bigarren gramatikak, aurreko gramatikaren etiketetan oinarrituta, aposizioan dagoen bigarren izen-sintagma markatzen du, betiere aposizio-sintagma izateko baldintzak betetzen baditu. Bi etiketak dituen sintagma aposizio-sintagma izango da. Gramatiketan jasotzen diren erregelak idazteko erabili ditugun ezaugarri linguistikoak kategoriaren, azpikategoriaren eta entitateen etiketak izan dira. Erregelak entitate motaren arabera (lekua, pertsona, erakundea edo bestelakoak) sailkatu ditugu, eta guztira 58 erregela daude: 37 lehenengo gramatikan eta 21 bigarrenean.

Lehenengo gramatikak ipintzen dituen erregelen adibide bat 4.16 irudian ikus dezakegu. Erregela horrek adierazten du:

- Esleitu]APOS1 etiketa (MAP),
- pertsona-entitate konposatu baten amaiera (ENTI_BUK_PER) adierazten duen hitz batean (TARGET),
- betiere (IF)
  - aurreko hitza (-1) pertsona entitate konposatu baten hasiera bada (ENTI_HAS_PER);
  - uneko hitzaren (O) hurrengo hitzak @KM>[27] funtzio sintaktikoa (NEXT_KM) badu;
  - hurrengo hitza (1) izen arrunta (IZE + ARR) bada,

[27] Funtzio sintaktiko horretan ">" ikurrak adierazten du hurrengo hitzak daramala funtzio sintaktiko nagusietako bat.

– eta ez (NOT) badu @KM> funtzio sintaktikoa (NEXT_KM).

Mota horretako erregelekin aposizio-sintagmak izateko hautagaiak eta lehen sintagmak etiketatzen dira.

```
MAP (]APOS1) TARGET (ENTI_BUK_PER) IF (-1 ENTI_HAS_PER)  (0 NEXT_KM)
                                     (1  IZE + ARR) (NOT 1 NEXT_KM);
```

4.16 irudia – Aposizio-sintagma izateko hautagaia etiketatzen duen erregela bat

4.17 irudian ikus daiteke aposizioen bigarren sintagma etiketatzeko eta aposizioa konfirmatzeko bigarren gramatikan biltzen den erregeletako bat. Erregela horrek adierazten du:

- Esleitu [APOS2 etiketa (MAP)
- izen (IZE) batean (TARGET),
- baldintza hauek (IF) betetzen badira:
  - uneko hitzak derrigorrez (OC) arrunta izan behar du (hau da, izen arrunta),
  - ez da (NOT 0) puntu kardinal (PUNT_KARDI) bat edo (OR) toki-postposizio (LEKU) bat izango,
  - ez da (NOT 0) hasiera hizki larriz izango (HM) edo (OR) dena hizki larriz (DM) egongo;
  - aurreko hitza (-1) aposizio hautagaia izan behar da (lehendabiziko gramatikak jarritako etiketa izan behar du);
  - eta hurrengo hitza (1) ez (NOT) da adjektiboa (ADJ) izango.

```
MAP ([APOS2) TARGET  (IZE) IF (0C ARR) (NOT 0 PUNT_KARDI OR LEKU )
                                (NOT 0 HM OR DM) (-1 APO) (NOT 1 ADJ) ;
```

4.17 irudia – Aposizio-sintagmaren bigarren izen-sintagma etiketatzen eta aposizio-sintagma egiaztatzen duen erregela bat

Erregela horiek (56) adibideko aposizioen mugak detektatuko dituzte eta motaren arabera sailkatuko dituzte.

(56) Luis Uranga]APOS1 presidenteak[APOS2 (...)

Gramatika horietan definitu ditugun etiketak 4.5 taulan aurkezten ditugu. Etiketa horien bidez aposizio-mota ezberdinen sailkapena[28] egin dugu eta sailkapen horrekin zein sinplifikazio-erregela aplikatu behar den errazago jakin dezakegu.

Aposizio-mota	1. osagaia	2. osagaia
1A	]APOS1	[APOS2
1B	]APOS1_KONTRA	[APOS2_KONTRA
2	]APOS1SINT	[APOS2SINT

4.5 taula – Gramatikek aplikatzen dituzten etiketak

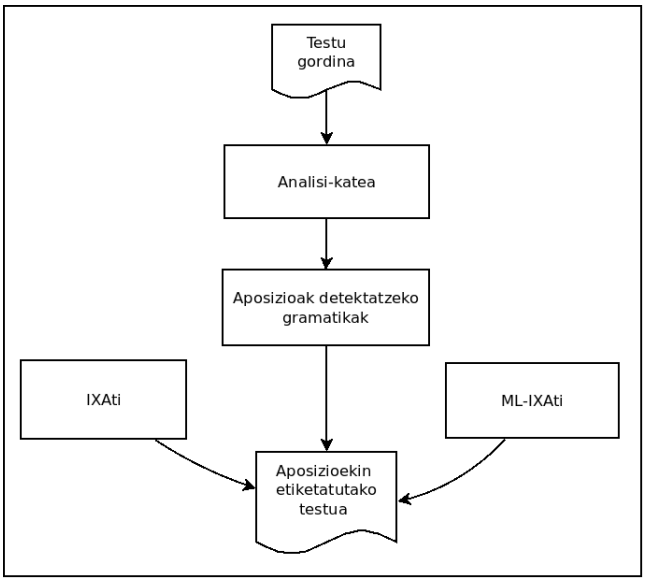
Behin aposizioa etiketatuta dagoela, erregeletan oinarritutako Ixati (Aduriz *et al.*, 2004) zatitzailearen eta ikasketa automatikoan oinarritutako ML-Ixati (Arrieta, 2010) zatitzailearen konbinazioa aplikatu dugu. Zehazki, aipamen-detektatzailean (Soraluze *et al.*, 2012) inplementatutako bertsioa erabili dugu. Aposizio-sintagman dauden bi izen-sintagmak osorik lortzeko, heuristiko hauek aplikatu ditugu: 1) lehenengo izen-sintagma hasiko da zatitzaileak zati-hasiera etiketatu duen hitzean, eta bukatuko da gure lehen gramatikak etiketa jarri duen hitzean. 2) Bigarren izen-sintagma lehenengo eta bigarren etiketaren artean dauden hitzek osatuko dute. Gramatiken eta zatitzailearen konbinazioak osatzen dute aposizio-detektatzailea (4.18 irudia).

Aposizio-detektatzailea garatzeko eta ebaluatzeko EPEC corpusa erabili dugu. Zehazki, MuGa gramatika garatzeko eta ebaluatzeko erabili dugun zati eta banaketa bera erabili dugu kasu honetan ere. Gramatika ebaluatzeko aposizioen urre-patroia eraiki dugu, eta bertan aposizio-sintagmak eskuz etiketatu ditugu. 4.6 taulan corpusean aurkitutako aposizio-kopurua (motaren arabera ere sailkatuta) eta detekzioan lortutako emaitzak[29] aurkezten ditugu.

Emaitzetan ikusten dugun bezala, mota guztientzat lortzen dugun F-neurria egokia da (0,80). Mota gehienetan onargarria da, baina 2 motakoetan baxuagoa da (0,62), estaldura dezente jaisten delako (0,44). Izan ere,

[28] 1A eta 1B motak izen-sintagma baten barnean gertatzen diren aposizio-sintagmak dira, 1A-n entitatea agertzen da aposizioan dagoen lehenengo sintagman eta 1B-n entitatea aposizioan dagoen bigarren sintagman agertzen da. 2 mota izen-sintagma osoak aposatuz egiten den aposizio-mota da.

[29] Erabilitako neurrien formulak: Doitasuna = zuzen detektatutako aposizio-sintagmak / detektatutako aposizio-sintagmak; estaldura = zuzen detektatutako aposizio-sintagmak / aposizio-sintagma guztiak; F-neurria = 2 * doitasuna * estaldura / (doitasuna + estaldura)



4.18 irudia – Aposizio-detektatzailearen arkitektura

	Kopurua	**Doitasuna**	**Estaldura**	***F*-neurria**
Guztira	336	0,87	0,74	0,80
1A mota	286	0,90	0,62	0,73
1B mota	30	0,85	0,73	0,79
2 mota	9	1	0,44	0,62

4.6 taula – Aposizio-detektatzailearen emaitzak

estaldura da gramatika horiek duten arazo nagusia, doitasuna, oro har, oso altua delako.

Detekzio-emaitzak kualitatiboki analizatu ditugu, eta hauek dira errore-analisiaren ondorioak: a) entitateen detekzioan izandako erroreengatik erregelak ez dira aplikatu edo gaizki aplikatu dira; b) aposizioa detektatu da, baina etiketa ez zegoen toki zuzenean (hau da, adibidez, adjektiboan etiketatu beharrean izenean jarri du etiketa); eta c) garapen-prozesuan jada alde batera utzitako aposizio konplexuak (aposizio koordinatuak, esaterako), horien detekzioak errore gehiago zekartzatelako.

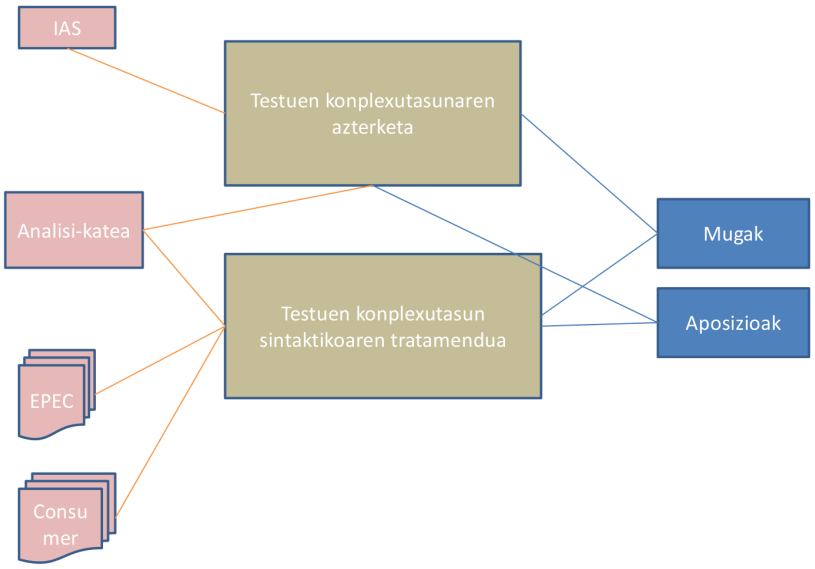
Sailkapenari dagokionez, kasu batean izan ezik aposizio guztiak zuzen sailkatu dira. Errore hori 1A motan sailkatu den egitura parentetiko bat izan da.

4.4 Laburpena

Kapitulu honetan, euskarazko testuen konplexutasuna automatikoki tratatzeko testuak sinplifikatuko dituen EuTS sistemaren oinarri linguistikoak aurkeztu ditugu. Sistema hori erregeletan (ezagutza linguistikoan) oinarrituta egongo da eta sintaxian bi sinplifikazio-mota egingo ditu: ordezkapen sintaktikoen sinplifikazioa eta sinplifikazio sintaktikoa. Testua sinplifikatu behar den eta zein sinplifikazio-mailaren arabera sinplifikatu behar den jakiteko, sinplifikazio-erabakien algoritmoa diseinatu dugu. Algoritmo horrek jatorrizko testuaren konplexutasunaren arabera erabakitzen du testua sinplifikatu behar den ala ez, eta testuaren hartzailearen araberako sinplifikazio-maila hautatzen du. Sinplifikazio sintaktikoa egiteko, sinplifikazio sintaktikoaren prozesua definitu dugu. Erregela horiek ere zein hurrenkeratan aplikatuko diren izan dugu hemen aztergai, eta dependentzia-zuhaitzean goitik behera agertu ahalako fenomenoen erregelak aplikatzea erabaki dugu. Horretaz gain, erregelen aplikazioan murriztapenak ezarri, eta txertatze-elementu alternatiboen erabilera zehaztu dugu. Etorkizunean integratu nahi ditugun ezaugarriak ere aipatu ditugu.

Bestetik, testuen konplexutasunaren azterketa eta sinplifikazio automatikoa egin aurretik beharrezkoa den testuen analisi automatikoa egiteko tresnak aurkeztu ditugu. Horien artean, Ixa taldearen analisi-katean erabiltzen diren tresnak aurkeztu ditugu, eta testuen konplexutasuna automatikoki analizatzeko eta testuen sinplifikazioa automatikoki gauzatzeko beharrezkoak diren Mugak eta Aposizioak tresnak hobetu, garatu eta ebaluatu ditugu. Tresna horiek txosten honen sarreran aurkeztutako irudiari gehitu dizkiegu, eta 4.19 irudian ikus daitezke.

Testuen konplexutasuna automatikoki analizatuko duen ErreXail sistema 5. kapituluan aurkeztuko dugu, eta testuak automatikoki sinplifikatuko dituen EuTS sistema 6. kapituluan.



4.19 irudia – Tesian erabilitako baliabideak eta tresnak, eta egindako ekarpenak

5 Konplexutasunaren analisi automatikoa: ErreXail sistema

Kapitulu honetan konplexutasunaren analisi automatikoa egiten duen ErreXail sistemaren arkitektura deskribatuko dugu. ErreXail inplementatzeko erabili ditugun ezaugarri linguistikoak aurkeztuko ditugu eta ikasketa automatikoarekin egindako esperimentuen berri emango dugu. Halaber, ErreXail beste hizkuntzetako sistemekin alderatu dugu.

5.1 Sarrera

Konplexutasunaren analisi automatikoa HPan *Redability Assessment* izeneko ikerketa-lerroak egiten du. Ikerketa-lerro hori oso zabala da, eta gure helburua testuen sinplifikazio automatikoa izanik, soilik gure helburu bera duten lanak izan ditugu kontuan.

ErreXail ezaugarri linguistikoetan oinarritzen den eta ikasketa automatikoko teknikak erabiltzen dituen sistema da. Sistema hori gai da testu bat zaila edo erraza den erabakitzeko. Hori dela-eta, ErreXail testuen sinplifikazio automatikoan aurreprozesu bezala erabiliko dugu, testuak sinplifikatu behar diren ala ez jakiteko.

ErreXail sistemak badu aitzindari bat euskararen prozesamenduan: Idazlanen Autoebaluaziorako Sistema (IAS) (Castro-Castro *et al.*, 2008; Aldabe *et al.*, 2012). Sistema horren helburua ikasleek egin dituzten testuei nota bat ematea da eta, horretarako, irakasleek testuak zuzentzeko dituzten irizpideak

jasotzen ditu: perpaus-kopurua esaldiko, esaldi-motak, perpaus-motak, kategoria ezberdinak eta lema-kopurua. Hasiera batean aurreprozesu moduan IAS erabiltzea pentsatu bagenuen ere, erdaretan egindako lanak ikusi ondoren hankamotz geratzen zitzaigula ikusi genuen. Gabezia horri aurre egiteko, erabaki genuen ErreXail sortzea.

5.2 Ezaugarri linguistikoak

ErreXailek, testuen konplexutasuna erabakitzeko, 94 ezaugarri linguistiko hartzen ditu kontuan. Ezaugarri horiek mailakatuta daude: orokorrak, lexikalak morfologikoak, morfosintaktikoak, sintaktikoak eta pragmatikoak, eta bakoitza analisi-katearen maila batekin lotuta dago. Ezaugarri horiekin testuaren monitorizazio linguistikoa egiten da.

Ezaugarri linguistikoen analisiaren emaitza (monitorizazio linguistikoa) kopuruetan eta ratioetan ematen da. Kopuruekin ratioak kalkulatzen dira; ratio horiek ikasketa automatikoan erabiltzen dira. Bestela esanda, testuen ezaugarriak ateratzen dira analisi horren ostean. Ondoren zerrendatuko ditugun ezaugarrietako batzuk jada beste hizkuntzetan erabili izan dira, baina beste batzuk euskararako bereziki sortu ditugu. Ezaugarri horiek guztiak inplementatzeko Perl programazio-lengoaia erabili dugu. ErreXailek egiten duen monitorizazio linguistikoaren adibide zati bat 5.1 irudian ikus daiteke.

```
KOPURUAK:
Hitz kopurua: 189
Karaktere kopurua: 936
Esaldi kopurua: 11
Perpaus kopurua: 38
Aposizio kopurua: 0
Kategoriak:
        Izen kopurua: 63
                Izen berezi kopurua: 0
                Leku-izen berezi kopurua: 0
        Aditz kopurua: 62
                Aditz laguntzaile kopurua: 38
                Aditz-izen kopurua: 3
                Aditz modal eta semi modal kopurua: 0
                Aditz faktitibo kopurua: 0
                Nor aditzen kopurua: 25
                Nor-nork aditzen kopurua: 10
                Nor-nori aditzen kopurua: 1
                Nor-nori-nork aditzen kopurua: 0
                Geroaldiko aspektua duten aditzen kopurua: 0
                Aspektu burutua duten aditzen kopurua: 0
                Aspektu ez burutua duten aditzen kopurua: 11
                Aspektu puntukaria duten aditzen kopurua (aditz trikoak): 30
                Orainaldiko aditzen kopurua: 35
                Lehenaldiko aditzen kopurua: 0
                Alegialdiko aditzen kopurua: 0
                Geroaldi arkaikoa (suposotiboa) duten aditzen kopurua: 0
                Indikatiboko aditzen kopurua: 33
                Subjuntiboko aditzen kopurua: 0

RATIOAK:
 Testu honi dagozkion ratioak ondorengo hauek dira.
Orokorrak:
Esaldien batazbesteko hitz kopurua 17.1818181818182 da.
Esaldien batazbesteko mendeko perpaus kopurua 3.45454545454545 da.
Hitzen batazbesteko karaktere kopurua 4.95238095238095 da.
Kategorien emaitzak:
Izenen eta hitz guztien arteko ratioa 0.333333333333333 da.
        Izen berezien eta izen arrunten arteko ratioa 0 da.
        Entitateen eta izen arrunten arteko ratioa 0.0158730158730159 da.
Aditzen eta hitz guztien arteko ratioa 0.328042328042328 da.
        Aditz-izenen eta aditz guztien arteko ratioa 0.0483870967741935 da.
        Aditz modalen eta aditz guztien arteko ratioa 0 da.
        Aditz faktitiboen eta aditz guztien arteko ratioa 0 da.
        Nor aditzen eta aditz guztien arteko ratioa 0.403225806451613 da.
        Nor-nork aditzen eta aditz guztien arteko ratioa 0.161290322580645 da.
        Nor-nori aditzen eta aditz guztien arteko ratioa 0.0161290322580645 da.
        Nor-nori-nork aditzen eta aditz guztien arteko ratioa 0 da.
        Geroaldiko aspektua duten aditzen eta aditz guztien arteko ratioa 0 da.
        Aspektu burutua duten aditzen eta aditz guztien arteko ratioa 0 da.
        Aspektu ez burutua duten aditzen eta aditz guztien arteko ratioa 0.17741935483871 da.
        Aspektu puntukaria duten aditzen eta aditz guztien arteko ratioa 0.483870967741935 da.
        Orainaldiko aditzen eta aditz guztien arteko ratioa 0.564516129032258 da.
        Lehenaldiko aditzen eta aditz guztien arteko ratioa 0 da.
        Alegialdiko aditzen eta aditz guztien arteko ratioa 0 da.
        Geroaldi arkaikoa (suposotiboa) duten aditzen eta aditz guztien arteko ratioa 0 da.
        Indikatiboko aditzen eta aditz guztien arteko ratioa 0.532258064516129 da.
        Subjuntiboko aditzen eta aditz guztien arteko ratioa 0 da.
        Ahalerazko aditzen eta aditz guztien arteko ratioa 0.0645161290322581 da.
```

5.1 irudia – ErreXailen monitorizazio linguistikoaren irteera

5.2.1 Ezaugarri orokorrak

Ezaugarri orokorrek testu osoa hartzen dute kontuan eta testuaren azaleko begiratu bat eskaintzen dute. Ezaugarri horiek 5.1 taulan ikus daitezke.

Batezbestekoak
Hitzen batezbestekoa esaldi bakoitzeko
Perpausen batezbestekoa esaldi bakoitzeko
Hizkien batezbestekoa hitz bakoitzeko

5.1 taula – Ezaugarri orokorren zerrenda

Ezaugarri horiek konplexutasuna neurtzeko formula klasikoetan oinarritzen dira. Lehenengo biak IASen kontuan hartzen dira.

5.2.2 Ezaugarri lexikalak

Ezaugarri lexikalak lemetan oinarritzen dira, hau da, analisiko lemen informazioa dute aztergai. Kategoria, sigla, laburdura eta sinbolo guztien ratioak kalkulatzen dira. Halaber, izen eta aditz-mota ezberdinak kontuan hartzen dira. Talde honetan guztira 39 ratio daude, eta horien laburpena 5.2 taulan ikus daiteke.

Ratioak
Behin agertzen diren lemak / lema guztiak
Kategoria bakoitza / hitz guztiak
Izen bereziak / izen guztiak
Entitateak / izen guztiak
Aditz-izenak / aditz guztiak
Aditz modalak / aditz guztiak
Aditz kausatiboak / aditz guztiak
Nor aditzak / aditz guztiak
Nor-Nori aditzak / aditz guztiak
Nor-Nork aditzak / aditz guztiak
Nor-Nori-Nork aditzak / aditz guztiak
Siglak / hitz guztiak
Laburdurak / hitz guztiak
Sinboloak / hitz guztiak

5.2 taula – Ezaugarri lexikalen zerrenda

Euskararako bereziki sortu ditugun ezaugarri lexikalak aditz arazleen ra-

tioa eta aditz-mota ezberdinen ratioak dira. Kategoria-mota ezberdinen ratioa IASen jasotzen zen.

5.2.3 Ezaugarri morfologikoak

Ezaugarri morfologikoek lemek har ditzaketen formak hartzen dituzte kontuan eta testuan dauden forma ezberdinak analizatzea dute helburu. Guztira 24 ratio daude eta 5.3 taulan zerrendatu ditugu.

Ratioak
Kasu-marka bakoitza / kasu-marka guztiak
Aditz-aspektu bakoitza / aditz guztiak
Aditz-aldi bakoitza / aditz guztiak
Aditz-modu bakoitza / aditz guztiak
Elipsia duten hitzak / hitz guztiak
Elipsi-mota bakoitza / elipsia duten hitzak

5.3 taula – Ezaugarri morfologikoak

Kontuan hartzen diren ezaugarriak honako hauek dira: kasu-marketan, analisi-kateko 18 kasu-markak (absolutiboa, ergatiboa, inesiboa, adlatiboa, genitiboa...); aspektuan, puntuala edo aoristoa, burutua, ez-burutua eta gerokoa; aldian, orainaldia, lehenaldia, irreala eta geroaldi arkaikoa; eta moduan, indikatiboa, subjuntiboa, inperatiboa eta potentziala. Elipsiari dagokionez, hitz-mailan gertatzen den elipsia da neurtzen dena, adibidez *dioguna* bezalako kasuak.

5.2.4 Ezaugarri morfosintaktikoak

Ezaugarri morfosintaktikoak azaleko sintaxian (zatietan edo *chunk*etan) eta aposizio-sintagmetan oinarritzen dira. Ratio horiek 5.4 taulan jaso ditugu.

Ratioak
Izen-sintagmak (*chunkak*) / sintagma guztiak
Izen-sintagmak (*chunkak*) / esaldi guztiak
Aditz-sintagmak / sintagma guztiak
Aposizio-sintagmak / sintagma guztiak
Aposizio-sintagmak / izen-sintagma guztiak (*chunkak*)

5.4 taula – Ezaugarri morfosintaktikoen zerrenda

Orain arte aurkeztu ditugun ezaugarrietan ez bezala, ezaugarri morfosintaktikoek hitz bat baino gehiago izan dezakete kontuan. Aposizio-sintagmei dagokienez, aposizio-detektatzaileak (ikus 4.3.3 atala) ematen dituen bi motak batera hartzen ditugu ratio horietan.

5.2.5 Ezaugarri sintaktikoak

Ezaugarri sintaktikoek mendeko perpaus-mota ezberdinen eta mendeko perpausen batezbestekoak hartzen dituzte kontuan. Guztira 10 ratio bildu ditugu eta 5.5 taulan laburbildu ditugu.

Ratioak
Mendeko perpausak / perpaus guztiak
Perpaus erlatiboak / mendeko perpausak
Perpaus osagarriak / mendeko perpausak
Perpaus adberbialak / mendeko perpausak
Perpaus adberbial mota bakoitza / mendeko perpausak

5.5 taula – Ezaugarri sintaktikoen zerrenda

Lan honetan aztertu ditugun perpaus adberbialak analisi-katean agertzen direnak dira: denbora-, kausa-, baldintza-, modu-, kontzesio-, ondorio- eta modu/denbora-perpausak. Azken horiek *-ela* menderagailuarekin eraikitzen diren eta modua eta denbora adierazten duten perpausak dira.

Hasierako lan honetan dependentzia-zuhaitzetan oinarritutako ezaugarriak (dependentzia-zuhaitzen sakonera, mendekotik gobernatzailerako distantzia...) ez erabiltzea erabaki dugu. Batetik, dependentzien analisi automatikoak denbora-kostua dauka, eta prozesua motelduko luke. Bestetik, sintaxiaren erabilera eztabaidatzen da literaturan. Petersen eta Ostendorfek (2009) esaten dute sintaxiak ez duela eragin handirik, eta Sjöholm-ek (2012) erakusten du dependentzia-zuhaitzak ez direla beharrezkoak. Pitler eta Nenkova-k (2008), aldiz, sintaxiaren garrantzia nabarmentzen dute. Dena den, Dell'Orletta *et al.*ek (2011) frogatu dute testu-mailan emaitza onak lor daitezkeela sintaxirik gabe, baina beharrezkoa dela esaldi-mailako sailkapenerako.

5.2.6 Ezaugarri pragmatikoak

Lan honetan ezaugarri pragmatikoek kohesio-mekanismoak hartzen dituzte kontuan. 12 ratio daude guztira eta 5.6 taulan laburtu ditugu.

Ratioak
Juntagailu-mota bakoitza / juntagailu guztiak
Lokailu-mota bakoitza / lokailu guztiak

5.6 taula – Ezaugarri pragmatikoen zerrenda

Juntagailu-motak emendiozkoak, hautakariak eta aurkaritzakoak dira. Lokailu-motak, berriz, emendiozkoak, hautakariak, aurkaritzakoak, azalpenezkoak, kausazkoak, ondoriozkoak, kontzesiozkoak eta modalak dira.

5.3 Ikasketa automatikoarekin egindako esperimentuak eta emaitzak

Behin testuaren ezaugarri linguistikoak analizatu ditugula, bi esperimentu egin ditugu ikasketa automatikoko teknikak aplikatuz. Lehenengo esperimentuan, testuak konplexuak edo sinpleak diren esango digun sailkatzaile bat eraikitzeko proba ezberdinak egin ditugu. Bigarren esperimentuan, ataza horretan ezaugarri esanguratsuenak zeintzuk diren jakin nahi izan dugu. Bi esperimentu horiek egiteko WEKA tresna (Hall *et al.*, 2009) erabili dugu.

Esperimentuak egiteko, bi corpus sortu ditugu saretik jaitsitako testuekin[1]. Lehenengo corpusa osatzeko, *Elhuyar* aldizkaritik[2] 200 dokumentu hartu ditugu (100 erreportaje eta 100 albiste) eta *Elhuyar (T-comp)* corpusa deitu dugu. Dibulgazio zientifikoko testuak dira eta corpus konplexutzat hartu dugu. Bigarren corpusa osatzeko, *ZerNola* webgunetik[3] beste 200 testu jaitsi ditugu eta *Zernola (T-simp)* corpusa deitu dugu. *ZerNola* webgunean Elhuyar Fundazioak osatutako eta haurren artean zientzia ezagutarazteko testuak daude, eta corpus sinpletzat hartu dugu. Bertatik jaso ditugun testu guztiak artikuluak dira. Corpusen kopuru nagusiak 5.7 taulan ikus daitezke.

Testu konplexuen eta testu sinpleen corpusak osatzeko, helduen eta haurren testuak alemanez ere (Hancke *et al.*, 2012) bildu dituzte.

[1]Elhuyar Fundazioko Lorea Arakistainen eta Iñaki San Vicenteren baimenari eta laguntzari esker eskuratu ditugu testu guztiak.

[2]http://aldizkaria.elhuyar.org/ (2014ko urtarrilean atzituta)

[3]http://www.zernola.net/ (2014ko urtarrilean atzituta)

Corpusa	Testuak	Esaldiak	Tokenak	Aditzak	Izenak
T-comp	200	8.593	161.161	52.229	59.510
T-simp	200	2.363	39.565	12.203	13.447

5.7 taula – *Elhuyar (T-comp)* eta *Zernola (T-simp)* corpusen ezaugarri nagusiak

5.3.1 Sailkatzailea eraikitzen

Lehendabiziko esperimentuan, testuak sinpleak edo konplexuak diren erabakiko duen sailkatzaile bat sortzea dugu helburu. Horretarako, bost sailkatzaile probatu eta ebaluatu ditugu. Sailkatzaile horiek ausazko zuhaitzak (*random forest*) (Breiman, 2001), J48 erabaki-zuhaitza (Quinlan, 1993), K-Nearest Neighbour (KNN) familiako IBk (Aha *et al.*, 1991), Naïve Bayes (John eta Langley, 1995) eta euskarri-bektoredun makinen SMO algoritmoa (Platt, 1998) dira. Antzeko beste lanetan bezala, ebaluazioa egiteko 10 geruzako balidazio gurutzatua erabili dugu.

Ezaugarri linguistiko guztiak batera kontuan hartuta, emaitza onenak SMOrekin lortu ditugu. Instantzien % 89,50 zuzen sailkatu ditu; testu konplexuen F neurria[4] % 0,899 izan da eta sinpleena % 0,891; MAE neurria[5], berriz, % 0,105. Algoritmo guztiekin lortutako zuzen sailkatutako dokumentuen emaitzak 5.8 taulan jaso ditugu.

Random Forest	J48	IBk	Naïve Bayes	SMO
88,50	84,75	72,00	84,50	**89,50**

5.8 taula – Ezaugarri guztiak erabiliz lortutako sailkapen-emaitzak

Ezaugarri-mota bakoitza ere bere aldetik probatu dugu. Soilik ezaugarri lexikalak erabiliz, emaitza hobea lortu dugu; hain zuzen ere, dokumentuen % 90,75 zuzen sailkatzea lortu dugu SMO sailkatzailea erabilita. Gainontzeko motekin lortutako emaitzak 5.9 taulan aurkeztu ditugu, eta soilik sailkatzaile onenek emandakoak adierazi ditugu.

Sailkatzaileak	Random Forest	J48	SMO
Orokorrak	74,25	73,50	**74,75**

(Jarraipena hurrengo orrialdean)

[4] F-neurria= 2 * doitasuna * estaldura / (doitasuna + estaldura)

[5] MAE neurriak, ingelesez *mean absolute error*, erroreen magnitudeen batezbestekoa adierazten du.

Sailkatzaileak	Random Forest	J48	SMO
Lexikalak	88,00	85,00	**90,75**
Morfologikoak	**82,00**	71,75	75,00
Morfosintaktikoak	**78,25**	76,25	72,75
Sintaktikoak	71,25	**73,75**	67,75
Pragmatikoak	67,50	**70,50**	65,75

5.9 taula – Ezaugarri-motaren arabera lortutako sailkapen-emaitzak

Ezaugarri-motak ere multzokatu ditugu konbinazio ezberdinak probatuz. SMO sailkatzailea erabilita, ezaugarri lexikalekin, morfologikoekin, morfosintaktikoekin eta sintaktikoekin osatutako multzoarekin hobetu ditugu orain arte erakutsitako emaitzak, dokumentuen % 93,50 zuzen sailkatuz. Multzokatze horretan lortutako emaitzarik onenak 5.10 taulan erakutsi ditugu.

Ezaugarri-multzoa	Random Forest	SMO
Oro+Lex	87,50	**89,50**
Oro+Lex+Morf	87,75	**89,00**
Oro+Lex+Morf+Morfsint	89,25	**89,50**
Oro+Lex+Morf+Morfsint+Sint	87,25	**90,25**
Morf+Morfsint	**84,25**	82,25
Morf+Morfsint+Sint	**83,25**	80,75
Morf+Morfsint+Sint+Prag	**83,75**	82,00
Lex+Morf	88,75	**92,75**
Lex+Morf+Morfsint	**89,25**	**89,25**
Lex+Morf+Morfsint+Sint	89,75	**93,50**
Lex+Morf+Morfsint+Sint+Prag	88,50	**90,25**
Sint+Prag	**78,25**	73,50

5.10 taula – Ezaugarri-motak multzokatuz lortutako sailkapen-emaitzak

Ezaugarriak multzokatuz, SMO da sailkatzailerik onena, baina ausazko zuhaitzek emaitza hobetu dute ezaugarri lexikalik egon ez denean. Esperimentu honetan ikusi ahal izan dugu, testuen konplexutasuna aztertzean, dokumentu-mailan ari garenean behintzat, ezaugarri lexikalek berebiziko garrantzia dutela, eta sintaxia, aldiz, ez dela hain erabakigarria.

5.3.2 Ezaugarri esanguratsuenak aztertzen

Ikasketa automatikoarekin egindako bigarren esperimentuan, testu konplexuak eta sinpleak bereizten dituzten ezaugarri esanguratsuenak zein diren

zehaztu nahi izan dugu. Ezaugarrien rankinga osatzeko, WEKAren *Information Gain (InfoGain AttributeEval)* erabili dugu. Bi ranking probatu ditugu: i) ezaugarri guztiekin egindako rankinga eta ii) ezaugarri-moten araberako rankinga. 5.11 taulan ezaugarri guztiak kontuan hartuta gehien erabiltzen diren 10 ratioak eta euren esanguratsutasuna aurkeztu ditugu.

Ratioa	**Esanguratsutasuna**
Izen berezien eta izen arrunten arteko ratioa (Lex.)	0,2744
Aposizio-sintagmen eta izen-sintagmen arteko ratioa (Morfosint.)	0,2529
Aposizio-sintagmen eta sintagmen arteko ratioa (Morfosint.)	0,2529
Entitateen eta izen arrunten arteko ratioa (Lex.)	0,2436
Behin bakarrik agertzen diren lemen eta lema guztien arteko ratioa (Lex.)	0,2394
Siglen eta hitz guztien arteko ratioa (Lex.)	0,2376
Aditz faktitiboen eta aditz guztien arteko ratioa (Lex.)	0,2099
Modu/denbora-perpausen eta mendeko perpaus guztien arteko ratioa (Sint.)	0,2056
Destinatiboaren eta kasu-marka guztien arteko ratioa (Morf.)	0,1968
Azalpenezko lokailuen eta lokailu guztien arteko ratioa (Prag.)	0,1957

5.11 taula – Gehien iragartzen duten 10 ratioen zerrenda eta bakoitzaren esanguratsutasuna

Ezaugarri-moten araberako rankingak ere osatu ditugu. 5.12 taulan ezaugarri orokorrak eta lexikalak zerrendatu ditugu; 5.13 taulan morfologikoak, eta morfosintaktikoak eta 5.13 taulan sintaktiko eta pragmatikoak.

Orokorrak	**Lexikalak**
Hitzen batez besteko karaktere-kopurua	Izen berezien eta izen arrunten arteko ratioa
Esaldien batez besteko hitz-kopurua	Entitateen eta izen arrunten arteko ratioa
	Behin bakarrik agertzen diren lemen eta lema guztien arteko ratioa
	Siglen eta hitz guztien arteko ratioa
	Aditz faktitiboen eta aditz guztien arteko ratioa

5.12 taula – Ezaugarri orokor eta lexikal esanguratsuenak

Morfologikoak	**Morfosintaktikoak**
Destinatiboaren eta kasu-marka guztien arteko ratioa	Aposizio-sintagmen eta izen-sintagmen arteko ratioa
Leku denborazko genitiboaren eta kasu-marka guztien arteko ratioa	Aposizio-sintagmen eta sintagmen arteko ratioa
Deskribatzailearen eta kasu-marka guztien arteko ratioa	Izen-sintagmen eta sintagmen arteko ratioa
Motibatiboaren eta kasu-marka guztien arteko ratioa	Aditz-sintagmen eta sintagmen arteko ratioa
Adlatiboaren eta kasu-marka guztien arteko ratioa	Izen-sintagmen eta esaldien arteko ratioa

5.13 taula – Ezaugarri morfologiko eta morfosintaktiko esanguratsuenak

Sintaktikoak	**Pragmatikoak**
Modu/denbora-perpausen eta mendeko perpaus guztien arteko ratioa	Azalpenezko lokailuen eta lokailu guztien arteko ratioa
Kontzesio-perpausen eta mendeko perpaus guztien arteko ratioa	Ondoriozko lokailuen eta lokailu guztien arteko ratioa
Baldintza-perpausen eta mendeko perpaus guztien arteko ratioa	Elipsidun aditzen eta elipsidun hitz guztien arteko ratioa
Denbora-perpausen eta mendeko perpaus guztien arteko ratioa	Kausazko lokailuen eta lokailu guztien arteko ratioa
Kausa-perpausen eta mendeko perpaus guztien arteko ratioa	Elipsidun hitzen eta hitz guztien arteko ratio

5.14 taula – Ezaugarri sintaktiko eta pragmatiko esanguratsuenak

Esperimentu horien emaitzak oso interesgarriak dira testuen sinplifikazio automatikorako. Izan ere, egitura konplexuak estatistikoki zeintzuk diren jakin dugu eta emaitzek fenomeno horiek lantzearen beharra erakusten dute. Esanguratsutasunik ez duten ezaugarriak ere analizatu ditugu; eta horietako batzuk dira, adibidez, inesiboaren eta gainontzeko kasu-marken arteko ratioa, indikatiboaren eta gainontzeko aditz-moduen ratioa, adjektiboen eta hitz guztien arteko ratioa, eta orainaldiaren eta gainontzeko aditz-aldien ratioa.

Amaitzeko, sailkapen-esperimentua (lehenengo esperimentua) errepikatu dugu bigarren esperimentu honetan lortu ditugun 10 ezaugarri esanguratsuenak erabiliz (ikus 5.11 taula). Sailkapen honen emaitzak 5.15 taulan aurkezten ditugu.

10 ezaugarri garrantzitsuenak erabilita dokumentuen % 88,25 zuzen sailkatzea lortu dugu, J48 sailkatzailearekin. Hori izan da, halaber, J48 sailka-

Random Forest	J48	IBk	Naïve Bayes	SMO
87,75	**88,25**	72,00	83,25	87,00

5.15 taula – Sailkapen-emaitzak 10 ezaugarri esanguratsuenak erabiliz

tzailearen emaitzarik onena.

Ikasketa automatikoko esperimentuak laburbilduz, testu-mailako konplexutasunaren analisian, emaitzarik onenak ezaugarri lexikalen, morfologikoen, morfosintaktikoen eta sintaktikoen multzoarekin eta SMO sailkatzailearekin lortu ditugu (dokumentuen % 93,50 zuzen sailkatuta). Ezaugarri lexikalen garrantzia ere azpimarratu nahi dugu, soilik horiek erabilita lortzen den emaitza ezaugarri guztiak erabilitakoa baino hobea baita, eta horietariko bost 10 ezaugarri esanguratsuenen artean aurkitu ditugu.

5.4 Testuen sinplifikazio automatikoa helburu duten erdaretako sistemak

Lehenik eta behin aipatu nahi dugu ezin dela konparazio zehatzik egin, beste hizkuntzetako tresnekin erabilitako datu-multzoak ezberdinak direlako. Hala ere, ErreXailek beste hizkuntzetako tresnen antzerako emaitzak lortzen dituen ala ez jakiteko egin dugu lan hau. Dena den, bakarrik gure helburu bera, hau da, testuak sinplifikatu behar ote diren jakitea duten lanekin alderatu dugu.

Konplexutasunaren araberako testuen sailkapena beste lanetan egin da aurretik. Gure corpusaren tamaina txikia izanik, soilik testu konplexuen eta sinpleen arteko bereizketa egiteko gai izan gara, Dell'Orletta *et al.*ek (2011) eta Hancke *et al.*ek (2012) egin duten bezala. Beste lan batzuetan (Schwarm eta Ostendorf, 2005; Aluísio *et al.*, 2010; François eta Fairon, 2012) konplexutasun-maila gehiago sailkatzeko gai izan dira. Aztergai izan ditugun lan guztietan, guk egiten dugun bezala, lehenik ezaugarri linguistikoak analizatzen dituzte eta ondoren ikasketa automatikoko teknikak aplikatzen dituzte.

Konparazioa gure antzeko teknikak erabiltzen dituzten lanekin hasiko dugu. Alemanerako, Hancke *et al.*ek (2012) bi konplexutasun-maila sailkatzen dituzte eta SMO erabiltzen dute, 10 geruzako balidazio gurutzatuarekin eba-

luatuz. Ezaugarri guztiak erabiliz, dokumentuen % 89,70 zuzen sailkatzea lortu dute, gure emaitzetatik gertu. Hala ere, kontuan izan behar da beraien corpusak 4.603 dokumentu dituela eta gureak, aldiz, 400. Ezaugarriak banan-banan probatuta, emaitzarik onena morfologikoekin lortu dute (% 85,40 zuzen sailkatuta), eta ezaugarri lexikalak, hizkuntza-ereduak eta ezaugarri morfologikoak konbinatuta dokumentuen % 89,40 zuzen sailkatu dituzte. Ezaugarri esanguratsuenak zein diren ezagutzeko, hauek ere *Information Gain* erabiltzen dute, baina ez dugu ezaugarri komunik.

Italierarako, Dell'Orletta *et al.*ek (2011) hiru esperimentu egin dituzte konplexutasuna neurtzeko, baina bakarrik lehenengoa dator bat gure lanarekin. Esperimentu horretarako 638 dokumentu erabili dituzte, eta bektoreen arteko distantzia euklidearra da erabili duten teknika. 5 geruzako balidazio gurutzatuarekin ebaluatu dute. Ezaugarri guztiak erabiliz, dokumentuen % 97,02 zuzen sailkatzea lortu dute, eta ezaugarri gordinak, lexikalak eta morfosintaktikoak multzokatuta, % 98,12.

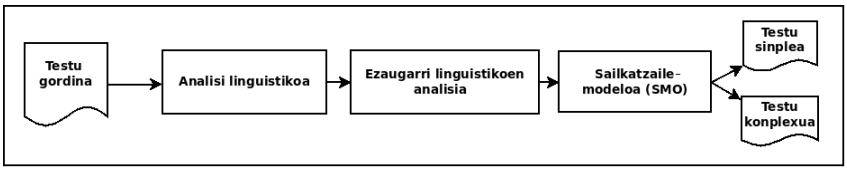
Brasilgo portugeserako, Aluísio *et al.*ek (2010) hiru mailatan sailkatu dituzte testuak: funtsezkoa, oinarrizkoa eta aurreratua. Guztira 592 testu jaso dituzte, eta SMO, 10 geruzako balidazio gurutzatua eta sailkapen estandarra erabiliz, 0,276 MAE tasa lortzen dute ezaugarri guztiak erabiliz. Jatorrizko testuen F neurria 0,913 da, sinplifikazio naturalean 0,483 eta sinplifikazio absolutuan 0,732. Ezaugarri-motekin esperimentuak egin dituzte, eta emaitzarik onenak ezaugarri guztiak erabiliz lortu dituzte. Korrelazio altua duten ezaugarrien artean, guk bezala, aposizio-sintagmen intzidentzia markatu dute bigarren postuan. Ez dugu beste ezaugarri komunik, ordea.

Ingeleserako, Feng *et al.*ek (2010) LIBSVM (Chang eta Lin, 2001), eta WEKAko erregresio logistikoa erabili dute, 10 geruzako balidazio gurutzatuarekin ebaluatuz. Testu mailakatuak erabili dituzte, eta LIBSVMrekin dokumentuen % 59,63 zuzen sailkatzea lortzen dute, eta erregresio logistikoarekin % 57,59. Maila ezberdinak kontuan hartzen dituztenez eta sailkatzaile ezberdinak erabiltzen dituztenez, ezinezkoa zaigu emaitzak alderatzea, baina ezaugarri esanguratsu batzuk amankomunean dauzkagu: entitateen dentsitatea eta izen bereziak.

5.5 ErreXail sistemaren arkitektura

Ezaugarri linguistikoak aztertuta eta ikasketa automatikoarekin esperimentuak egin eta gero, ErreXail sistemaren arkitektura aurkeztuko dugu. Erre-

Xailek hiru mailako arkitektura dauka eta 5.2 irudian ikus daiteke.



5.2 irudia – ErreXail sistemaren arkitektura

Beraz, euskarazko testu bat emanda, urrats hauek jarraituko ditugu:

1. Testuaren analisi linguistiko automatikoa. Analisi hori 4.3 atalean aurkeztutako tresnekin egiten da.

2. Testuaren monitorizazio linguistikoa. Neurketa hori egiteko, 5.2 atalean aurkeztutako ezaugarriak erabiltzen dira.

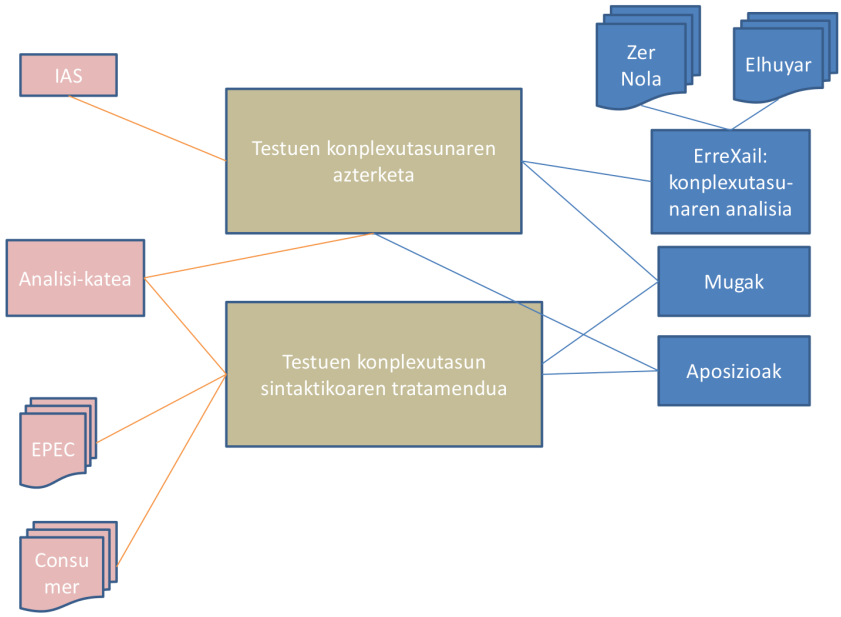
3. SMO Support Vector Machine sailkatzailearen erabilera. Sailkatzaile hori erabiltzen da emaitzarik onenak eman dituelako, 5.3 ataleko esperimentuetan azaldu bezala. Sinplifikazioaren aurreprozesua bizkortzeko, ezaugarri lexikalen, morfologikoen, morfosintaktikoen eta sintaktikoen multzokatzea erabiltzen da.

ErreXail testuen sinplifikazioaren aurreprozesua izateko jaio bada ere, ErreXail sistema guztiz autonomoa da eta beste helburuekin edo aplikazioekin erabil daiteke. Izan ere, ErreXail erabili da itzulpen automatikoen postedizioa egitean testu ezberdinen konplexutasuna neurtzeko (Aranberri *et al.*, 2014). ErreXail sistemari beste ezaugarri batzuk gehituta eta ikasketa automatikoko beste teknika batzuk erabilita, Madrazok (2014) B1, B2, C1 eta C2 mailetako testuak bereizten dituen sistema bat egin du.

ErreXail sistemak dituen ezaugarri linguistikoak beste hizkuntzetara ere eramangarriak dira, beti ere beste hizkuntzen analisiak ahalbidetzen duten bitartean. Teknika estatistikoak ere, noski, beste hizkuntzetan aplika daitezke. Etorkizunerako, beste hainbat ezaugarri inplementa daitezke: elkarketaren eta eratorpenaren analisia, lexikoaren maiztasunak edo dependentzia-zuhaitzen sakonera. Hizkuntza-ereduak ere erabil daitezke. Ezaugarri gehiago gehitzeaz gain, corpus handiagoak eta maila ezberdinetakoak lortuz gero, maila ezberdinak bereizteko ere molda daiteke.

5.6 Laburpena

Kapitulu honetan testuen konplexutasuna aztertzen duen ErreXail sistema aurkeztu dugu. ErreXailek hizkuntzaren fenomenoak kontuan hartzen dituen 94 ratio analizatzen ditu eta, ondoren, ikasketa automatikoko teknikak erabiliz testua sinplea ala konplexua den adierazten digu. Esperimentuetan ezaugarririk esanguratsuenak ere zeintzuk diren ikusi dugu. Sistema hori testuak sinplifikatu behar diren ala ez jakiteko erabiliko dugun arren, sistema autonomoa da. Kapitulu honetan aurkeztu dugun ErreXail sistema entrenatzeko, *Elhuyar (T-comp)* eta *Zernola (T-simp)* corpusak sortu ditugu. Kapituluaren ekarpenak 5.3 irudian gehitu ditugu.



5.3 irudia – Tesian erabilitako baliabideak eta tresnak, eta egindako ekarpenak

Behin testuak konplexuak ala sinpleak diren jakinda, testu konplexuak sinplifikatuko dituen EuTS sistema aurkeztuko dugu 6. kapituluan.

6

Euskarazko testuen sinplifikazio automatikoa: EuTS sistemaren diseinua

Kapitulu honetan analisi linguistikoan proposatutako sinplifikazioa garatuko duen EuTS sistemaren aurkezpena egingo dugu. EuTS sistemak bi motatako sinplifikazioak egingo ditu: ordezkapen sintaktikoen sinplifikazioa eta sinplifikazio sintaktikoa. Sinplifikazio-mota horiek egingo dituzten moduluen eta behar duten informazio linguistikoaren deskribapena egingo dugu hemen.

6.1 Sarrera

EuTS (Euskarazko Testuen Sinplifikatzailea) euskarazko testuen sinplifikazio automatikoa egingo duen sistema da. Tesi-lan honen helburu nagusietako bat izan da sistema honen diseinu teorikoa egitea eta sistemak behar dituen moduluen informazio linguistikoa garatzea edo egokitzea. Hau da, EuTS sistemaren inplementazioa ez dugu guztiz garatu, baina sistemaren moduluak informazio linguistikoaz hornitu ditugu. Dena den, kasu-azterketa bezala, egitura parentetikoei dagozkien erregelak inplementatu ditugu eta emaitza ebaluatu dugu.

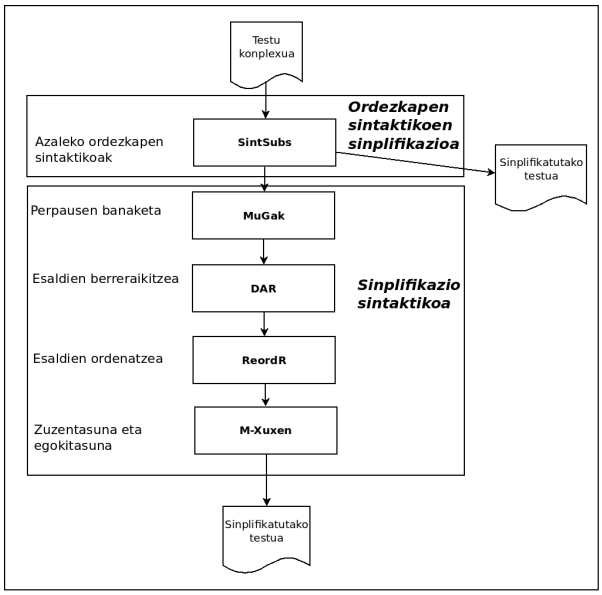
EuTS sistemak tesi-lan honetan egindako analisi linguistikoari jarraikiz, bi sinplifikazio-mota egingo ditu: i) maiztasunetan oinarritzen den ordezkapen sintaktikoen sinplifikazioa eta ii) egitura-aldaketak gauzatzen dituen sinplifikazio sintaktikoa. Bi sinplifikazio-mota horiek aurrera eramateko bost

eragiketa (azaleko ordezkapen sintaktikoak, banaketa, esaldien berreraikitzea, esaldien ordenatzea eta esaldien zuzenketa eta egokitzapena) egingo ditu. Eragiketa horiek sistemaren arkitekturan bost moduluetan egingo dira (ikus 6.1 irudia). Jarraian, modulu bakoitzak zein eragiketa egingo duen azaltzen dugu:

- **Ordezkapen sintaktikoen sinplifikazioa:**
  - **SintSubs modulua, azaleko ordezkapen sintaktikoak:** egitura sintaktikoen azaleko ordezkapena.
- **Sinplifikazio sintaktikoa:**
  1. **Mugak modulua, banaketa:** esaldiak perpausetan banatu, aposizioak aposizio-sintagmetan eta egitura parentetikoak perpaus nagusietan eta tartekietan.
  2. **DAR (***deletion and addition rules* **modulua, esaldien berreraikitzea:** ezaugarri morfologikoak (menderagailuak, kasu-markak...) eta bestelako elementuak (juntagailuak, postposizioak...) ezabatu eta esanahia mantenduko duten elementuak (adberbioak, izen-sintagmak...) txertatu; hau da, erlazio-markak ezabatu eta txertatze-elementuak txertatu.
  3. **ReordR modulua, esaldien ordenatzea:** esaldi berrien barneko elementuak ordenatu eta esaldi berriak euren artean ordenatu.
  4. **M-Xuxen modulua, esaldien zuzenketa eta egokitzapena:** akatsak zuzendu, testuaren kohesioa bermatu.

EuTS sistemak, eragiketa horiek aplikatu aurretik, bi aurreprozesu izango ditu: jasoko dituen sarrerako testuak automatikoki analizatutako dira 4.3 atalean aurkeztutako tresnekin. Behin testuen analisia daukagula, EuTS sistemara pasatu aurretik, sinplifikazio-erabakien algoritmoan ikusi bezala (4.2.1 atala), testu horien konplexutasuna ErreXail sistemak (5. kapitulua) aztertuko du. Ondoren, hartzailearen araberako sinplifikazio-mota eta sinplifikazio-mailak aukeratuko dira.

Erdal hizkuntzetako sistemen arkitekturekin erkatzen badugu, EuTS sisteman analisia eta konplexutasunaren ebaluazioa aurreprozesu bezala definitu ditugu eta ez sistemaren modulu bezala. SintSubs, Mugak eta DAR moduluak Siddharthan-en (2006) sistemako transformazio-moduluko parte izango lirateke, eta ReordR eta M-Xuxen moduluak, berriz, birsortze-modulukoak.



6.1 irudia – EuTS sistemaren arkitektura

6.2 Ordezkapen sintaktikoen sinplifikazioa

Atal honetan, EuTS sisteman ordezkapen sintaktikoen sinplifikazioa egiten duen **SintSubs modulua** (*syntactic subtitutions*) aurkeztuko dugu. Modulu honek egiten dituen eragiketak azaleko ordezkapen sintaktikoak egitea da eta testuak azaleko sinplifikazio sintaktikoa (ASS) izeneko sinplifikazio-mailara egokitzen ditu.

6.2.1 SintSubs modulua: azaleko ordezkapen sintaktikoak

Azaleko ordezkapen sintaktikoak (AOS) maiztasunetan oinarrituta egiten diren eragiketak dira. **SintSubs** moduluak AOSen bitartez egitura sintaktiko bat baliokidea den maiztasun handiagoko batekin ordezkatzen du, eta aditz ez-jokatua duten perpaus adberbialetan aplikatzen du.

Azaleko ordezkapen sintaktikoak definitzeko, EPEC-DEP corpusean (Aranzabe, 2008) egin dugun azterketan oinarritu gara, 3. kapituluan azaldu dugun bezala. Azterketa horretan, *Euskal Gramatika: Lehen Urratsak*

gramatikan (Euskaltzaindia, 1999, 2005, 2011) agertzen diren egitura sintaktikoen zerrenda osatu dugu (A eranskina), eta egitura horiek corpusean agertzen diren eta, agertuz gero, zenbatetan agertzen diren aztertu dugu, besteak beste.

Maiztasun gutxiko edo maiztasunik gabeko 39 egitura ordezkatzeko, 16 aukera definitu ditugu, eta aukera horiek inplementatzeko 42 erregela idatzi ditugu. Perpaus ez-jokatu mota bakoitzak dituen helburu-egiturak, ordezkapen-aukerak eta erregela-kopuruak 6.1 taulan zehaztu ditugu. Aipatu behar dugu ere ordezkapen-aukera ez dela sekula helburu-aukeretako bat.

	Helburu-egiturak	**Ordezkapen-aukerak**	**Erregela-kopurua**
Guztira	39	17	42
Denbora	15	5	16
Kausa	1	1	1
Helburua	5	1	5
Baldintza	6	3	9
Kontzesioa	2	1	3
Modua	10	6	8

6.1 taula – Egituren, ordezkapen-aukeren eta erregelen kopuruak

Erregelak aplikatzeko, adierazpen erregularretan oinarritutako Perl programa bat inplementatu dugu. Programa horrek testu hutsean egiten du lan, eta egitura ezezagunak (corpusean aurkitu ez direnak edo maiztasun gutxi daukatenak) bilatu eta maiztasun altuagoko baliokideekin ordezkatzen ditu. Adibidez, *-tzearren* egitura (57) *-tzeko* egiturarekin ordezkatzeko, 6.2 irudiko erregela aplikatzen du.

(57) a. *Abuztuaren amaieran beste goi bilera bat egitea aztertzen ari dira Israel eta PAN Palestinako Aginte Nazionala, Ekialde Erdiko bake prozesua suspertzearren.*

b. i. *Abuztuaren amaieran beste goi bilera bat egitea aztertzen ari dira Israel eta PAN Palestinako Aginte Nazionala, Ekialde Erdiko bake prozesua suspertzeko*.

Erregela horretan (6.2 irudia) adierazi dugu:

- Corpusean *-tearren* edo *-tzearren* egiturak[1] aurkitzen baldin baditu,
- *-arren* karaktere-katea *-ko* karaktere-katearekin ordezkatu

```
#  tzearren -> tzeko
if (($corpusa=~/.*tearren[ \.\,\;\:]/)
   || ($corpusa=~/.*tzearren[ \.\,\;\:]/)) {
          $corpusa =~ s/arren/ko/ ;
}
```

6.2 irudia – Ordezkapen sintaktikoak (helburu-perpaus ez-jokatu) egiten dituen erregela

Erregelei dagokienez, 6.2 irudian ikusi dugun bezala, erregelak perpaus ez-jokatuen oinarri ezberdinak (aditz-izena eta partizipioa) eta kasu-markak edo erlazio-markak hartzen ditu kontuan eta, ondoren, horiek ordezkatzen ditu. Egiturak eta ordezkapen-aukerak oinarri bera badute, hau da, biek oinarri bezala aditz-izena (*-tze, -te*) edo partizipioa (*ø, -tu, -du, -i*) badute, erlazio-marka ordezkatzen du (adibidean, *-arren -> -ko*). Oinarria aldatuz gero, oinarria ere aldatu egiten da, partizipiotik aditz-izenera aldatuz edo alderantziz.

SintSubs moduluko erregelak inplementatzeko, *EGLU DS* datu-multzoa bildu dugu. Datu-multzo horretan EPEC-DEPen agertzen ez diren egiturak jaso ditugu, eta, horretarako, Euskaltzaindiaren EGLU gramatikan agertzen diren adibideak erauzi ditugu. Une honetan aditz ez-jokatuak dituzten perpaus adberbialen adibideak (Euskaltzaindia, 2011) jaso ditugu; guztira 108 esaldi dira. Datu-multzo horretatik esaldien erdia erregelak inplementatzeko, hau da, programa diseinatzeko erabili dugu eta beste erdia testa egiteko. Zati bakoitzean 54 esaldi daude. Datu-multzo horren kopuruak 6.2 taulan ikus daitezke.

Zatia	Esaldiak	Perpausak	Hitzak
Entrenamendua	54	155	583
Ebaluazioa	54	138	588

6.2 taula – Esaldi-, perpaus- eta hitz-kopurua datu-multzoan

Ebaluazioa egiteko, bi parametro hartu ditugu kontuan: ordezkapen zuzena eta irteera gramatikalki zuzena (esaldi zuzena). Bigarren parametroa lehenengoa zuzena izan den kasuetan kalkulatu dugu. 6.3 taulan ordezkapen sintaktikoen sinplifikazioa egitean lortutako emaitzak aurkezten ditugu.

[1]Erregela horretan oinarria aditz-izena (*-te-* edo *-tze-*) da eta erlazio-marka *-arren*.

	Ordezkapen zuzenak	Esaldi zuzenak
Guztira	79,63	88,64
Denbora	62,50	93,34
Kausa	100,00	100,00
Helburu	100,00	100,00
Baldintza	88,89	62,50
Kontzesioa	100,00	100,00
Modua	90,00	100,00

6.3 taula – Ordezkapen sintaktikoen sinplifikazioaren emaitzak guztira eta motaz mota

Emaitzak aztertuta, ikus dezakegu perpaus-mota gehienen emaitzak onak eta onargarriak direla. Emaitzak okerrera doaz, aldiz, aldaketa gehiago eta egitura gehiago dituzten motetan (denbora eta baldintza). Errore-analisia egin dugu, eta erroreak kasu hauetan gertatzen dira: a) oinarria aldatzen den kasuetan (partizipio <-> aditz-izena) eta b) partizipioek ø marka dutenean. Lehenengo erroreak ordezkapen okerrak dakartza, eta, bigarrenak, berriz, esaldi ez-gramatikalak. Arazo horiei aurre egiteko, ordezkapenak analisimailan egin beharko lirateke morfologia sortzen duten tresnak (Alegria, 1995) eta Hizkuntzaren Sorkuntzaren teknikak (Agirrezabal *et al.*, 2015) erabiliz.

Maiztasunetan oinarritutako hurbilpena sinplifikazioari begira egokia dela ikusteko, galdeketa bidezko bi ebaluazio egin ditugu. Batetik, hizkuntzen irakaskuntzan eskarmentua duten bi hizkuntzalariri eta, bestetik, azaleko sinplifikazio sintaktikoa sinplifikazio-mailaren helburu-talde diren bi euskara ikasle aurreraturi eta maila altua ez duten bi hiztuni galdetu diegu ea sinplifikatutako esaldiak jatorrizkoak baino sinpleagoak iruditzen zaizkien. Ebaluazioa egin duten hizkuntzalariak Euskal Herriko lurralde ezberdinetakoak dira eta, bereziki, eskatu zaie beraien esperientzia eta ingurua kontuan izateko epaiak egiterakoan. Ebaluazioan parte hartu duten helburu-taldeko lau kideei beren balorazio propioa eskatu zaie. Esperimentuan parte hartu duten lau hiztunak Euskal Herriko lurralde ezberdinetakoak dira eta guztiek unibertsitateko ikasketak dituzte. Ebaluaziorako lagina zuzen ordezkatutako esaldiekin osatu dugu.

Hizkuntzalariek emandako iritziak 6.4 taulan jaso ditugu. Bertan agertzen diren ehunekoek esaldi-kopuruari egiten diote erreferentzia. Hau da, mendebaldeko hizkuntzalariaren arabera, sinplifikatutako esaldien % 76,74 sinpleagoa da jatorrizkoa baino eta % 23,26 antzekoa da. Ez zaio irudi-

tzen sinplifikatutako esaldirik jatorrizkoa baino konplexuagoa denik. Eki-erdialdeko hizkuntzalariak, aldiz, uste du sinplifikatutako esaldien % 30,23 jatorrizkoa baino sinpleagoa dela, % 48,84 antzekoa dela eta % 20,93 konplexuagoa dela.

	Sinpleagoa	**Antzekoa**	**Konplexuagoa**
Mendebaldeko hizkuntzalaria	76,74	23,26	0,00
Eki-erdialdeko hizkuntzalaria	30,23	48,84	20,93

6.4 taula – Sinpletasun-iritziak hizkuntzalarien arabera

Sortutako esaldien sinpletasuna ere, esan bezala, helburu-taldearekin aztertu dugu. Horretarako, maila aurreratua duten bi euskara ikasleri eta bi hiztun ez trebaturi eskatu diegu ea, beren iritziz, sinplifikatutako esaldiak sinpleagoak, antzekoak edo konplexuagoak diren. Emaitzak 6.5 taulan ikus daitezke. Laburbiltzeko asmoarekin, bertan ematen ditugun datuak oro har eman dituzten iritzien ehunekoak dira, esaldi guztiak eta perpaus adberbial mota ezberdinak kontuan hartuz.

	Sinpleagoa	**Antzekoa**	**Konplexuagoa**
Guztira	75,00	25,00	0,00
Denbora	33,33	58,33	8,33
Kausa	87,50	0,00	12,50
Helburua	75,00	25,00	0,00
Baldintza	25,00	0,00	75,00
Kontzesioa	37,50	37,50	25,00
Modua	75,00	25,00	0,00

6.5 taula – Helburu-taldeak egindako sinpletasun-iritziak

Ikus dezakegun bezala, esaldi guztiak kontuan hartuta, esaldien % 75,00 sinpleagoa iruditu zaie, eta gainontzeko % 25,00 antzekoa. Oro har, inorentzat ez dira konplexuago izan. Parte-hartzaileen jatorria kontuan izanda, eki-erdialdeko hiztunaren arabera esaldi gehienek antzeko konplexutasuna dute, eki-erdialdeko hizkuntzalariak esan bezala.

Emaitzak perpaus-motaren arabera aztertuta ikus dezakegu baldintza-perpausak oro har konplexuagoak direla sinplifikatu ondoren, eta denbora-perpausek jatorrizkoen antzeko zailtasuna dutela. Baina datu horiek jatorriaren arabera begiratuta, ikus dezakegu mendebaldeko hiztunaren ustez

baldintza- eta denbora-perpausak sinpleagoak direla sinplifikatu ondoren, mendebaldeko hizkuntzalariak aipatu bezala. Kontzesio-perpausetan ez dugu orokortzerik lortu eta gainontzeko perpausen emaitzak onak dira. Emaitza horiek ondorioztatzera eramaten gaituzte sinplifikazioaren helburu izango den pertsonaren jatorria ere (eta inguruko euskalkia) kontuan izan behar dela testuak sinplifikatzean.

6.3 Sinplifikazio sintaktikoa

Atal honetan EuTS sisteman sinplifikazio sintaktikoa egiten duten moduluak aurkeztuko ditugu. Modulu horiek dira Mugak, DAR (*Deletion and Addition Rules*), ReordR (*Reordering Rules*) eta M-Xuxen. Modulu horiek egiten dituzten eragiketak banaketa, esaldien berreraikitzea, esaldien ordenatzea eta esaldien zuzenketa eta egokitzapena dira. Modulu bakoitzak zer egingo duen adierazteko 4. kapituluan azaldutako (45) adibidearen analisia (6.3 irudia) baliatuko dugu, irudi horretan agertzen diren etiketak baitira moduluek tratatuko dituztenak.

```
"<Asperren>"<HAS_MAI>"   S:137/0
        "asper" IZE ARR GEN NUMP MUGM ZERO HAS_MAI w1,L-A-IZE-ARR-10,lsfi2 @IZLG> %SIH S:137 %ESALDI_HAS_1_BEREZIA &NCMOD>
"<kasua>"
        "kasu" IZE ARR BIZ- ABS NUMS MUGM w2,L-A-IZE-ARR-14,lsfi3 @SUBJ %SIB &NCSUBJ>
"<emeki-emeki>"
        EZEZAG "emeki-emeki" ADB ARR ZERO w3,L-G-ADB-ARR-2,lsfi6 @ADLG %SINT &NCMOD>
"<aitzinatu>"
        "aitzinatu" ADI SIN PART BURU NOTDEK w4,L-A-ADI-SIN-8,lsfi7 @-JADNAG %ADIKATHAS &MENOS>
"<bada>"
        "izan" ADL BALD A1 NOR NR_HURA w5,L-A-ADL-3,lsfi8 @+JADLAG_MP_ADLG %ADIKATBU &MENOS>
"<ere>"
        "ere" LOT LOK EMEN w6,L-A-LOT-LOK-6,lsfi11 @LOK &CMOD>
"<,>"<PUNT_KOMA>"  S:608/0
       PUNT_KOMA S:608 }MUGA
"<Sa>"<HAS_MAI>"
        EZEZAG "Sa" IZE IZB PLU- ENTI_HAS_PER AORG HAS_MAI w8,L-G-IZE-IZB-3,lsfi12 @KM> %SIH &NCMOD>
"<Pintoren>"<HAS_MAI>"
        "pinto" ADJ ARR IZAUR- GEN MG ENTI_BUK_PER HAS_MAI w9,L-A-ADJ-ARR-10,lsfi14 @IZLG> &NCMOD>
"<etorkizuna>"
        "etorkizun" IZE ARR BIZ- ABS NUMS MUGM w10,L-A-IZE-ARR-18,lsfi15 @SUBJ> %SIB &NCSUBJ>
"<fite>"
        "fite" ADB ARR ZERO w11,L-A-ADB-ARR-3,lsfi18 @ADLG %SINT &NCMOD>
"<argituko>"
        "argitu" ADI SIN PART GERO NOTDEK w12,L-A-ADI-SIN-10,lsfi19 @-JADNAG %ADIKATHAS
"<da>"
        "izan" ADL A1 NOR NR_HURA w13,L-A-ADL-5,lsfi20 @+JADLAG %ADIKATBU &<AUXMOD
"<$.>"<PUNT_PUNT>"  S:123/0  S:148/0
       PUNT_PUNT S:123 %ESALDI_BUK_1 S:148 }MUGA
```

6.3 irudia – (45) adibideko esaldiaren analisi automatikoa

Sinplifikazio-mota honetan bi sinplifikazio-mailako testuak egokituko ditugu: sinplifikazio naturala eta sinplifikazio absolutua. Sinplifikazio natural mailako testuak lortuko ditugu perpaus koordinatuei, aditz jokatuak dituzten

mendeko perpausei, aposizio-sintagmei eta egitura parentetikoei dagozkien erregelak aplikatuta; sinplifikazio absolutukoak, berriz, horietaz gain, aditz ez-jokatuak dituzten mendeko perpausen eta landutako postposizio-egituren erregelak aplikatuta.

6.3.1 Mugak modulua: banaketa

Splitting edo banaketa **Mugak** moduluak egingo du. Modulu honen zeregin nagusia esaldian dauden perpausak banatzea da, baina horrez gain egitura parentetikoak, aposizio-sintagmak eta postposizio-egiturak banatuko ditu. Horretarako MuGa gramatikan, aposizio-detektatzailean eta postposizioen gramatikan oinarrituko da. Perpausak banatzeko luzera minimoa bete behar denez, luzera minimoa egiaztatzeko Ixatiren irteera baliatuko du.

Sinplifikazio sintaktikoaren prozesua abiatu aurretik, perpausen luzera minimoa betetzen dela egiaztatzeko, Ixatik emandako *chunk*ak zenbatuko ditu. Hortaz, perpaus bakoitzean %ADIKATBU (aditz-katearen bukaera, aditz perifrastikoekin) edo %ADIKAT (aditz-katearen bukaera, aditz trinkoekin) etiketez gain, %SINT (hitz bakarreko sintagma) eta %SIB (hitz anitzeko sintagmaren bukaera) etiketak bi aldiz daudela egiaztatuko du.

Aztergai dugun esaldian, etiketa horiek[2] lehenengo perpausean *kasua* eta *emeki-emeki* hitzen analisietan aurkitu ditugu, eta bigarrenean *etorkizuna* eta *fite* hitzenetan. Beraz, (45) adibideko esaldiak luzera minimoa betetzen duela egiaztatu dugu, osagai bat baino gehiagoko bi sintagmez, osagai bateko bi sintagmez eta bi aditz-katez osatutakoa delako.

Perpausak banatzeko, }MUGA etiketa baliatuko du. Horren aurretik agertzen dena perpaus bat izango da, eta perpausaren aurreko muga esaldi-hasierako etiketak (&ESALDI_HAS_1_BEREZIA edo &ESALDI_HAS_1)[3] edo beste perpaus-mugako etiketa bat izango dira.

Adibidean, }MUGA etiketa komari (7. tokena) eta amaierako puntuari (14. tokena) dagozkien analisietan aurkituko ditu; &ESALDI_HAS_1_BEREZIA etiketa, berriz, *Asperren* hitzean. Honenbestez, adibideko esaldia bi perpausetan banatuko du: 1) *Asperren kasua emeki-emeki aitzinatu bada ere,* eta 2) *Sa Pintoren etorkizuna fite argituko da.*

[2] Adibidean ikusten ditugun %SIH eta %ADIKATHAS etiketek sintagma-hasiera eta aditz-kate hasiera adierazten dute hurrenez hurren.

[3] &ESALDI_HAS_1_BEREZIA etiketa testuko lehen esaldiari jartzen zaio, eta testu- eta esaldi-hasiera adierazten du. &ESALDI_HAS_1 testuko lehen esaldia ez den gainontzeko esaldien hasiera adierazten duen etiketa da.

Aposizio-sintagmak banatzeko, 4. kapituluan azaldu bezala, aposizio-detektatzailearen emaitza (Ixatiren eta aposizioen gramatikaren konbinazioa) erabiliko du. Aposizioan dagoen lehenengo sintagma %SIH (hitz anitzeko sintagma-hasiera) eta]APOS1 edo]APOS1_KONTRA etiketek mugatuko dute, eta bigarren sintagma, berriz,]APOS1 edo]APOS1_KONTRA eta [APOS2 edo [APOS2_KONTRA etiketek. Izen-sintagma osoak aposatuz egiten diren aposizioetan, aposizioan dagoen sintagmaren hasierako muga ezagutzeko]APOS1SINT etiketa erabiliko du, eta amaierako muga ezagutzeko, aldiz, [APOS2SINT etiketa.

Egitura parentetikoak banatzeko, 6.4 atalean azalduko dugun Biografix tresna erabiliko du. Landu ditugun postposizio-sintagmak banatzeko, Ixatik ematen digun {POS-BUK etiketa baliatuko du.

6.3.2 DAR (*deletion and addition rules*) modulua: esaldien berreraikitzea

Esaldiak banatu ondoren, **DAR** (*Deletion and Addition Rules*) moduluak esaldiak berreraikiko ditu. Bi azpieragiketa nagusi egingo ditu: a) ezabatzea (*removing*) eta txertatzea (*adding*). Oro har, lehenengoan erlazio-markak mendeko perpausetik ezabatuko ditu (Erlazio_Marken_Zerrenda), eta, bigarrenean, ezabatutako esanahia berreskuratuko duten elementuak (Txertatze_Elementuen_Zerrenda) txertatuko ditu. Txertatze-elementuen aukeraketa ere hemen egingo da.

Modulu horrek Morfeusen eta Eustaggerren irteerak baliatuko ditu ezabatzeak inplementatzeko. Beharrezkoak izango dituen etiketa gehienak mendeko perpausen aditzen analisiei dagozkie, hau da, mendeko perpausen aditzen analisietan egin beharko ditu ezabatzeak. Etiketa horiek perpaus-mota adierazten dutenak eta funtzio sintaktikoak adierazten dutenak dira. Aditz jokatua duten mendeko perpausetan ezabatu behar diren ezaugarriak 6.6 taulan laburbildu ditugu. “Ezabatzeko ezaugarri morfologikoak” zutabean ezaugarri morfologikoak eta perpaus-motak adierazten dituzten etiketak ageri dira. Ezabatzeko bestelako elementuak dira, erlazio-markak konposatua direnean, menderagailuekin edo kasu-markekin batera dauden beste elementuak, esaterako *ba... ere, nahiz eta... den* eta *-en bezala*; normalean hitz bat edo hitz anitzeko terminoak dira. Kasu horietan morfema den elementua analisitik ezabatuko da eta beste elementua, normalean hitz funtzional bat dena, guztiz ezabatuko da. Funtzio sintaktikoetan (FS) ezabatu behar diren ezaugarriak

azpimarratuta daudenak dira, MP (mendeko perpausa) eta betetzen duten funtzioa: IZLG> = izenlaguna; SUBJ = subjektua; PRED = predikatiboa; OBJ = objektua; eta ADLG = adizlaguna.

Fenomenoa	**Perpaus-mota**	**Ezabatzeko ezaugarri morfologikoa**	**Ezabatzeko bestelako elementuak**	**FSetan ezabatzeko ezaugarriak**
Erlatiboak	Arrunta	ERLT		@+JADNAG_MP_IZLG>
	Zein erlatiboa	ZHG	*zein, non*	@+JADNAG_MP_IZLG>
	Zein erlatiboa	KAUS	*zein, non*	@+JADNAG_MP_IZLG>
Osagarriak	Konpletiboa	KONPL		@+JADNAG_MP_SUBJ
	Konpletiboa	KONPL		@+JADNAG_MP_PRED
	Konpletiboa	KONPL		@+JADNAG_MP_OBJ
	Zehar-galdera	ZHG	*ea*, galdetzaileak (NOLGAL azpi-kategoria)	@+JADNAG_MP_OBJ
	Zehar-galdera	ZHG	*ea*, galdetzaileak (NOLGAL azpi-kategoria)	@+JADNAG_MP_SUBJ
Adberbialak	Denbora	DENB		@+JADNAG_MP_ADLG
	Denbora	MOD/DENB		@+JADNAG_MP_ADLG
	Denbora	ZHG	*gehienetan, aldiro, bezain laster, bezain ber...*	@+JADNAG_MP_OBJ
	Denbora	ERLT	*guztietan..*	@+JADNAG_MP_IZLG>
	Kausa	KAUS		@+JADNAG_MP_ADLG
	Kontzesioa	MOS	*nahiz eta, arren*	@+JADNAG_MP_ADLG
	Kontzesioa	BALD	*ere*	@+JADNAG_MP_ADLG
	Modua	MOD/DENB		@+JADNAG_MP_ADLG
	Modua	MOS	*bezala, moduan...*	@+JADNAG_MP_ADLG
	Helburua	HELB		@+JADNAG_MP_ADLG
	Baldintza	BALD	*baldin*	@+JADNAG_MP_ADLG

6.6 taula – Aditz jokatuak dituzten mendeko perpausetan ezabatu behar diren ezaugarriak

Adibidera bueltatuz, *bada* hitzaren analisitik perpaus-mota adierazten duen BALD etiketa ezabatuko du, eta funtzio sintaktikoa adierazten duen @+JADLAG_MP_ADLG etiketari mendeko perpausa (MP) dela eta adizlagun (ADLG) funtzioa duela adierazten duten atalak ezabatuko dizkio,

@+JADLAG etiketa (aditz jokatu laguntzailea) lortzeko. Adibidean *ba- ere* egitura lantzen ari garenez, *ere* hitzari dagokion sarrera eta analisia ere ezabatuko ditu.

Txertatzeak egiteko, berriz, erregelaren arabera, dagokion posizioan txertatuko ditu txertatze-elementua eta horri dagokion analisia. Txertatuko elementuak (Txertatze_Elementuen_Zerrenda), 3. kapituluan aurkeztu ditugu eta B eranskinean jaso ditugu. Hizpide dugun adibidean, kontzesio-perpausa sinplifikatzen ari garenez, *Hala ere* dagokion analisiarekin txertatuko du perpaus nagusian:

```
"<Hala_ere>"   S:130/0
      <Correct!> "hala_ere" LOT LOK AURK mw1,L-A-LOT-LOK-3,lsfi1 @LOK S:130 &ESALDI_HAS_1
```

Txertatzeak morfologia-mailakoak badira, adibidez aspektu-markak aldatzea helburu-perpausen sinplifikazio-proposamenetan ikusi dugun bezala, ezabatze-eragiketan egin bezala, analisian egingo ditu.

Halaber, modulu honek bihurtuko ditu aditz ez-jokatuak aditz jokatu. Horretarako, aditzen eta perpausaren analisiaren informazioa baliatuko du: aditz ez-jokatutik argumentuen kasua, mendeko perpausetik komunztadura adierazten duten pertsonak, eta perpaus nagusiaren aditzetik aldia eta aspektua. Egun, eginkizun hori bigarren urrats aurreratu batekoa dela iruditzen zaigu, kasu batzuetan ere zaila baita eskuz zehaztea. Beraz, mementoz aditz ez-jokatuekin egingo duen sinplifikazio automatikoa ordezkapen sintaktikoen sinplifikazioa izango da.

Modulu hau inplementatzeko, Morfeus euskarazko morfologia-analizatzailearen sortzailea erabiliko dugu (Alegria, 1995). Erregelak idazteko, FOMA (Hulden, 2009) egoera finituko teknologia erabiltzea aurreikusi dugu.

6.3.3 ReordR modulua: esaldien ordenatzea

Esaldi berrien barneko elementuak ordenatzeaz eta esaldi berriak euren artean ordenatzeaz arduratuko den modulua **ReordR** (*Reordering Rules*) da. Esaldien hurrenkera berria ezartzeko, bi eragiketa ezberdin egingo ditu. Alde batetik, esaldi berrien barneko osagaien hurrenkera zehaztu behar du (adibidez aposizioetan eta egitura parentetikoetan) eta, beste aldetik, esaldi berri horien arteko hurrenkera testuan definitu behar du.

Esaldi barneko osagaien hurrenkera dela-eta, hasierako proposamenean jatorrizko esaldian dagoen hurrenkerari eutsiko dio, eta aposizioetatik eta egitura parentetikoen tartekietatik sortutako esaldien barne-hurrenkera zehaztuko du modulu honek.

Esaldi berrien arteko hurrenkerari dagokionez, 3. kapituluan aurkeztutako corpus-analisia erabili dugu erregelak osatzerakoan oinarri gisa (Hurrenkeren_Zerrenda). Beraz, azterketa horretatik ateratako ondorioak eta erabakiak aplikatuko ditu. Adibidera itzuliz, esaldiek jada mendekoa$_{jat}$-nagusia$_{jat}$ hurrenkera betetzen dutenez, modulu honek ez du aldaketarik egingo.

6.3.4 M-Xuxen modulua: esaldien zuzenketa eta egokitzapena

Testua berreraikita eta ordenatuta dagoenean, esaldien zuzentasuna **M-Xuxen** moduluak egiaztatuko du. Hasiera batean soilik Xuxen zuzentzaile ortografikoak (Agirre *et al.*, 1992) jatorrizko esaldiek izan ditzaketen balizko akatsak zuzenduko edo estandarizatuko ditu. Sinplifikatutako esaldi berrien puntuazio-markak ere modulu honek landuko ditu.

Beraz, adibidera itzuliz, modulu honek lehenengo esaldian amaierako puntua dagokion analisiarekin gehituko du:

```
"<$.>"<PUNT_PUNT>"  S:123/0  S:148/0
      PUNT_PUNT S:123 &ESALDI_BUK_1 S:148 }MUGA
```

Eta *Hala ere* lokailuaren ondoren, koma idatzi behar denez, koma ere dagokion analisiarekin gehituko du:

```
"<,>"<PUNT_KOMA>"
     PUNT_KOMA
```

Etorkizunean XuxenG zuzentzaile gramatikala ere integratuko genuke, esaldi berriak sortzean gerta daitezkeen akats gramatikalak zuzentzeko. Korreferentziaren azterketa (Soraluze *et al.*, 2015) ere integratuko dugu kohesioa bermatzeko. Honela, testuaren egokitasuna egingo genuke.

Lau modulu horien emaitza (45) adibideko esaldi sinplifikatuak izango dira. Esaldi sinplifikatu horien analisia 6.4 irudian ikus daiteke.

6.4 Kasu-azterketa: egitura parentetiko biografikoen sinplifikazio sintaktikoa

Atal honetan, 3. kapituluan, zehazki 3.3.6 atalean, deskribatutako egitura parentetikoen sinplifikazioaren kasu-azterketa aurkeztuko dugu. Hau da, egi-

```
"<Asperren>"<HAS_MAI>"   S:137/0
        "asper" IZE ARR GEN NUMP MUGM ZERO HAS_MAI w1,L-A-IZE-ARR-10,lsfi2 @IZLG> %SIH S:137 %ESALDI_HAS_1_BEREZIA &NCMOD>
"<kasua>"
        "kasu" IZE ARR BIZ- ABS NUMS MUGM w2,L-A-IZE-ARR-14,lsfi3 @SUBJ %SIB &NCSUBJ>
"<emeki-emeki>"
        EZEZAG "emeki-emeki" ADB ARR ZERO w3,L-G-ADB-ARR-2,lsfi6 @ADLG %SINT &NCMOD>
"<aitzinatu>"
        "aitzinatu" ADI SIN PART BURU NOTDEK w4,L-A-ADI-SIN-8,lsfi7 @-JADNAG %ADIKATHAS
"<da>"
        "izan" ADL A1 NOR NR_HURA w5,L-A-ADL-3,lsfi8 @+JADLAG %ADIKATBU &<AUXMOD
"<$.>"<PUNT_PUNT>"  S:123/0  S:148/0
       PUNT_PUNT S:123 &ESALDI_BUK_1 S:148 }MUGA
"<Hala_ere>"   S:130/0
        <Correct!> "hala_ere" LOT LOK AURK mw1,L-A-LOT-LOK-3,lsfi1 @LOK S:130 &ESALDI_HAS_1 &LOTAT
"<,>"<PUNT_KOMA>"
       PUNT_KOMA
"<Sa>"<HAS_MAI>"
        EZEZAG "Sa" IZE IZB PLU- ENTI_HAS_PER AORG HAS_MAI w4,L-G-IZE-IZB-3,lsfi12 @KM> %SIH &NCMOD>
"<Pintoren>"<HAS_MAI>"
        "pinto" ADJ ARR IZAUR- GEN MG ENTI_BUK_PER HAS_MAI w5,L-A-ADJ-ARR-10,lsfi14 @IZLG> &NCMOD>
"<etorkizuna>"
        "etorkizun" IZE ARR BIZ- ABS NUMS MUGM w6,L-A-IZE-ARR-18,lsfi15 @SUBJ %SIB &NCSUBJ>
"<fite>"
        "fite" ADB ARR ZERO w7,L-A-ADB-ARR-3,lsfi18 @ADLG %SINT &NCMOD>
"<argituko>"
        "argitu" ADI SIN PART GERO NOTDEK w8,L-A-ADI-SIN-10,lsfi19 @-JADNAG %ADIKATHAS
"<da>"
        "izan" ADL A1 NOR NR_HURA w9,L-A-ADL-5,lsfi20 @+JADLAG %ADIKATBU &<AUXMOD
"<$.>"<PUNT_PUNT>"  S:123/0  S:148/0
       PUNT_PUNT S:123 %ESALDI_BUK_1 S:148 }MUGA
```

6.4 irudia – (45) adibideko esaldi sinplifikatuen analisi automatikoa

tura parentetikoen sinplifikazio-proposamenak erregela bihurtu ditugu, eta erregela horiek inplementatu ditugu. Kasu-azterketa horren bitartez, EuTS sistemaren sinplifikazio sintaktikorako arkitekturaren diseinua eta informazio biografikoa duten egitura parentetikoak sinplifikatzeko diseinatutako erregelak probatu ditugu. Kasu azterketa horretan, EuTS sisteman proposatzen den arkitekturari jarraikiz, Biografix tresna sortu dugu egitura parentetikoen sinplifikazio sintaktikoan kontzentratzeko.

6.4.1 Biografix

Biografix patroietan oinarritutako tresna bat da, eta, egitura parentetikoak testutik banatzeaz gain, egitura horiek informazio biografikoa (jaiotza- eta heriotza-datuak) baldin badute, egitura horiek sinplifikatu eta esaldi sinple berriak sortzen ditu. Biografixek definitu dugun sinplifikazio sintaktikoaren prozesua eta EuTS sistemaren moduluak jarraitzen ditu. Aipatu behar dugu patroietan oinarrituta dagoenez, euskaraz gain beste hizkuntza batzuk trata ditzakeela; hau da, Biografix euskararako diseinatu bada ere, berreraikitze-eragiketa beste zazpi hizkuntzatara egokitu da: frantsesa, alemana, gaztelania, katalana, galegoa, italiera eta portugesa.

Biografixek honako hau egiten du eragiketa bakoitzean:

- **Banaketa:** landuko dituen esaldien elementuak banatzea. Horretarako, hiru urrats ematen ditu:

  1. Egitura parentetikoa perpaus nagusitik banatu
  2. Adierazpen parentetikoan datu biografikorik ba ote dagoen egiaztatu
  3. Datak eta tokiak banatu

  Patroi sinpleak erabiltzen ditu datak eta tokiak banatzeko; toki bat baino gehiago dagoen kasuetan, koma da banatzeko erabiltzen den marka. Eragiketa hau hizkuntza guztietan berdina den arren, alemaneko bertsioan bizitza- eta heriotza- datuak banatzeko erabiltzen den marka ezberdina denez (gidoia erabili beharrean puntu eta koma darabilte), kasu jakin hori ere egokitu dugu.

- **Berreraikitzea:** esaldi sinpleak sortzea. Lau esaldi-mota sortzen ditu:

  1. Perpaus nagusia egitura parentetikorik gabe
  2. Jaiotzaren informazioa duten esaldiak
  3. Heriotzaren informazioa duten esaldiak
  4. Toki-xehetasunak dituzten esaldiak

  Hizkuntza guztietarako lau esaldi-mota horiek sortzen baditu ere, ñabardurak daude esaldien sorkuntzan hizkuntza bakoitzaren ezaugarriak kontuan hartuta: aditzak, preposizioak, kasu-markak... Euskarako bertsioan, absolutibo kasua ere ezabatzen du (dataren formatutik datorrena). Eragiketa hau da hizkuntza bakoitzarekiko espezifikoa den bakarra.

- **Esaldien hurrenkera:** esaldiak testuan ordenatzea:

  1. Esaldi nagusia
  2. Jaiotzari buruzko informazioa duen esaldia.
  3. Jaiotzaren toki-xehetasunak dituen esaldia (baldin badago eta behar adina)
  4. Heriotzari buruzko informazioa (hilda badago) duen esaldia.

5. Heriotzaren toki-xehetasunak dituen esaldia (baldin badago eta behar adina)

- **Zuzenketa:** esaldiak zuzentzea. Egingo dugun ebaluazioan Biografixen zuzentasuna ebaluatuko dugunez, ez dugu eragiketa hau inplementatu.

Euskarazko bertsioa garatzeko, *Wikipedia*ko egitura parentetikoak nola idatzi behar diren jasotzen duten gidalerroetako irizpideak inplementatu ditugu, eta *Wikipedia DS* datu-multzoa (*Wikipedia DS*[4]) osatu dugu, gidalerroekin bat ez datozen instantziak ere jaso ahal izateko. Esaldi horiek ausaz hartu ditugu. Beste hizkuntzetan ez dugu garapenik egin, soilik 3-5 esaldi erabili ditugu errorerik ematen ez duela konprobatzeko.

Beste hizkuntzei dagokienez, gure helburua ez da izan beste hizkuntzentzat tresna bat sortzea, euskarazkoa beste hizkuntzetara aplika daitekeen ikustea baizik. Hau da, euskararako diseinatutako erregelak ea beste hizkuntzetan baliagarriak diren, eta euskararako inplementazioa ere beste hizkuntzetarako berrerabilgarria den. Horregatik, bertsio horiek Ixa taldearen webgunean[5] jarri ditugu eskuragarri, komunitatean norbaitek landu nahi baditu. Izan ere, hobekuntzak egin daitezke berreraikitze-eragiketan. Hizkuntza bakoitzean lehenaldi mota erabiliena erabili dugu, frantsesez eta italieraz izan ezik; frantsesez lehenaldi ezagunena eta erabiliena *passé composé* den arren, horrek komunztadura eskatzen du subjektuaren eta aditzaren artean[6]. Arazo hori ekiditeko guk *passé simple* aldia erabili dugu, baina *passé composé* aldira moldatzea izan daiteke egin daitekeen hobekuntza bat. Italieraz ere arazo bera dugu, eta horregatik *passato remoto* erabili dugu *passato prossimo* erabili beharrean. *Reordering* eragiketan ez da aldaketarik egin behar, baina zuzenketa hizkuntza bakoitzari egokitu behar zaio.

6.4.2 Biografixen ebaluazioa

Biografix ebaluatzeko bi ebaluazio egin ditugu: eskuzko ebaluazioa (intrintsekoa) eta ebaluazio estrintsekoa. Azken hori soilik euskarazko bertsioarekin egin dugu eta Seneko (Lopez-Gazpio eta Maritxalar, 2013)[7] web-aplikazioa

[4] *Wikipedia*tik 50 esaldi hartu ditugu (1.378 hitz) garapenerako eta beste 30 esaldi (715 hitz) ebaluazioa egiteko.

[5] https://ixa.si.ehu.es/Ixa/Produktuak/1403535629 (2014ko abuztuan atzituta)

[6] e.g. *Cher est né**e** en Californie.*, vs. *Ernest Rutherford est né**ø** en Angleterre.*

[7] http://ixa2.si.ehu.es/seneko/ (2014ko martxoan atzituta)

erabili dugu. Seneko *chunk*etan oinarritzen den galdera-sortzailearen (Aldabe *et al.*, 2013) aplikazioa da.

Biografix ebaluatzeko, hau da, *Wikipedia DS*ko ebaluazio zatia osatzeko, *Wikipedia*ko 30 artikuluren lehen esaldiak jaso ditugu. Hau da erabili dugun metodoa:

1. CatScan V2.0β[8] erabiliz, *Euskal Wikipedia*ko biografien zerrenda bat osatu dugu.

2. Zerrenda hori ausaz berrordenatu dugu eta beste bat osatu dugu landutako 8 hizkuntzatan zeuden artikuluak aukeratzeko.

3. Bigarren zerrenda horretatik lehen 32 artikuluak aukeratu ditugu.

32 artikulu horietatik lehenengo biak eskuzko ebaluazioa egingo duten pertsonak entrenatzeko erabili ditugu eta gainontzekoak eskuzko ebaluaziorako eta estrintsekorako. Hau da, guztira 30 artikulu erabili dira ebaluazioan.

Biografixen eskuzko ebaluazioa

Eskuzko ebaluazioa sei hizkuntzatan egin dugu: euskara, katalana, frantsesa, galegoa, alemana eta gaztelania. 10 hizkuntzalarik ebaluazio-prozesuan parte hartu dute eta hiru alde ebaluatu dituzte: jatorrizko esaldiak (*JatTestua*), Biografixen errendimendua (*Prog*) eta esaldi berrien gramatikaltasuna (*Gram*). Guztira bai/ez erantzun posiblea zuten bederatzi galdera erantzun dituzte. Proposatzen dugun ebaluazio-metodo honek errore-analisia egitea eta tresnaren puntu ahulak identifikatzea errazten du. Honako hauek dira egin dizkiegun galderak:

Jatorrizko esaldiari buruzko galderak (*JatTestua*). *Wikipedia*ko jatorrizko esaldietan egitura parentetikoak dauden eta jatorrizko esaldiak zuzenak diren jakiteko ala ez eta hiru galdera erantzun behar izan dituzte:

1. Parentesien artean idatzitako egitura parentetikorik ba al dago?

2. Esaldia gramatikalki zuzena eta estandarra da?

3. Puntuazio-markak zuzenak dira?

[8] http://tools.wmflabs.org/catscan2/catscan2.php (2014ko martxoan atzituta)

Parentesien artean idatzitako egitura parentetikoak dituzten esaldiek landutakoen multzoa (*coverage*) osatuko dute, hau da, gure 30eko laginetik zenbat landu ditzakeen Biografixek. Jatorrizko esaldien gramatikaltasunari eta puntuazio-markei buruz galdetzea erabaki dugu, Aldabe *et al.*en (2013) lanean ikusi zelako jatorrizko esaldi asko ez zirela zuzenak eta horrek galdera-sortzaileen errendimendua jaisten zuen.

Biografixen errendimenduari buruzko galderak (*Prog*). Biografixek bere zeregina betetzen duen ikusteko (doitasuna), lau galdera hauek egin ditugu:

1. Egitura parentetikoak kendu dira?
2. Informazio guztia mantentzen da?
3. Jatorrizko esaldia kontuan izanda, informazio guztia zuzena da?
4. Informazio berririk ba al dago?

Bigarren eta hirugarren galdera beharrezkoak dira berreraikitze-eragiketan informazioa galdu edo aldatu den jakiteko. Laugarren galderarekin jakin nahi dugu ea, adibidez, esaldi batean izenaren itzulpenak gehitu eta datu biografiko bezala tratatu diren, edo bizirik dagoen pertsona bati heriotza-informazioa esleitu zaion.

Sortutako esaldi sinplifikatuen gramatikaltasunari buruzko galderak (*Gram*). Bi galdera formulatu ditugu esaldi berrien zuzentasuna ebaluatzeko. Izan ere, esaldiak zuzenak izateak berebiziko garrantzia du testua ulertzean. Galdera hauek sortutako esaldi bakoitzeko erantzun behar dituzte (doitasun gramatikala).

1. Esaldia gramatikalki zuzena eta estandarra da?
2. Puntuazio-markak zuzenak dira?

Galdera horiekin Biografixen irteera konprobatuko da esaldi zuzenak sortzen dituen ikusteko. Sinplifikatutako esaldietan akatsik balego, zuzenduak izango lirateke sinplifikazio-prozesuaren zuzenketa-eragiketan.

Biografix ebaluatzeko erabili diren hizkuntzak *Wikipedia*ko gidalerroetan[9] agertzen diren formatu-aukeren arabera aukeratu ditugu:

- Euskara: Biografix diseinatu den helburu-hizkuntza
- Katalana: euskarak duen formatu bera
- Alemana: formatu bera, baina bariazio txiki bat
- Gaztelania: formatu bera, baina beste formatu batzuk ere aukeran
- Frantsesa: formatu parentetikoen artean formatu bera aukeran, eta beste aukera batzuk
- Galegoa: formatua definitu gabe

Portugesa eta italiera ez ditugu ebaluatu beren formatuak jada katalanarekin eta gaztelaniarekin ebaluatu direlako, hurrenez hurren. Katalana eta galegoa izan ezik, gainontzeko hizkuntzetako laginak bi hizkuntzalarik ebaluatu dituzte. Katalanarena euskararen formatu bera denez, eta galegoak formatu definitu gabea duenez, lagin horietako bakoitza soilik hizkuntzalari batek ebaluatu ditu, nahasteak ekiditeko. Parte hartu duten hizkuntzalari guztiak natiboak edo maila altua dute ebaluatu duten hizkuntzan; denak Ixa taldekoak dira edo izan dira.

Eskuzko ebaluazioaren emaitzak 6.7 taulan erakusten ditugu. Emaitza horiek ondorengo neurrien arabera antolatu ditugu:

1. Landutakoak (*coverage*): laginaren tamainatik (30etik) Biografixek landu dituen esaldien ehunekoa, hau da, egitura parentetikoak dituzten esaldien ehunekoa.
2. Zuzentasuna (*correctness*): jatorrizko esaldi zuzenen (gramatika eta puntuazio-markak) ehunekoa (*JatTestua* galderak).
3. Estaldura (*recall*): sortutako esaldien eta jatorrizko esaldien informazioa kontuan izanda sortu beharko lituzkeen esaldien ratioa.
4. Doitasuna (*precision*): ondo prozesatutako esaldien (*Prog* galdera guztiei zuzen erantzundakoak) eta prozesatutako esaldien arteko ratioa (*precision at performance*).

[9] *Wikipedia* entziklopediek estilo-liburuak eta gidalerroak dituzte artikuluak modu uniformean idazteko.

5. Doitasun gramatikala (*grammatical precision*): zuzen (gramatika eta puntuazio-markak) sortutako esaldien eta guztira sortutako esaldien ratioa (*Gram* galderak).

Taularen azken-aurreneko zutabean κ (Cohenen kappa, Cohen, 1960) adostasun-koefizientea erakutsi dugu, eta azkenekoan ehuneko adostasuna (adostasun-proportzioa, %). Instantzia gutxi dauzkagunez, eta adostasun proportzioa oso altua denez, horrek κ baxuak ekartzen ditu. Hori dela-eta, adostasun hori ere azaltzen dugu hemen.

Hizkuntza	**Landutakoak** %	**Zuzentasuna** %	**Estaldura**	**Doitasuna**	**Doitasun gram.**	κ	%
Euskara	97,00	82,76	0,94	0,79	0,87	0,37	90,63
Katalana	93,33	98,21	0,77	0,53	0,78	-	-
Frantsesa	73,00	88,64	0,80	0,18	0,37	0,39	85,06
Galegoa	43,00	88,46	0,76	0,15	0,62	-	-
Alemana	100,00	100,00	0,78	0,60	0,78	-	100,00
Gaztelania	100,00	85,00	0,71	0,33	0,67	0,52	88,76

6.7 taula – Biografixen emaitzak ebaluatutako hizkuntzetan

Euskarazko emaitzak aztertuz, Biografix ia esaldi guztiak sortzeko gai dela ikusten dugu (estaldura: 0,94), eta esaldi horiek zuzenak direla (doitasun gramatikala: 0,87), informazio guztia mantentzen eta zuzen mantentzen arazotxoak dauden arren (doitasuna: 0,79). Jatorrizko esaldien zuzentasuna baxua (% 82,76) denez, Aldabe *et al.*i (2013) jarraituz, emaitzak birkalkulatu ditugu soilik jatorrizko esaldi zuzenak kontuan hartuz. Honela, estaldura 0,93 da, doitasuna 0,80 eta doitasun gramatikala 0,88. Ikus daitekeen bezala, emaitzak ez dira asko aldatu; izan ere, gure kasuan jatorrizko esaldiak soilik perpaus nagusian du eragina. Ebaluatzaileen adostasuna aztertzean ikusten dugu κ oso baxua dela (0,37), batez ere gramatika ebaluatzean izandako desadostasunengatik; hala ere, adostasun-proportzioa altua da (90,63).

Katalanaren kasuan, ikusten dugu Biografix ez dela gai jatorrizko esaldietan dagoen adina informazio sortzeko (estaldura: 0,77), eta joera hori hemendik aurrerako gainontzeko hizkuntzetan aurkituko dugu. Errendimenduko doitasuna ere baxua (0,53) da gehitutako eta galdutako informazioarengatik, baina doitasun gramatikala onargarria da (0,78). Gure ustez, katalanaren egokitzapena egokitzapen txukuna izan da.

Frantseseko emaitzek, aldiz, zerbait gaizki dagoela adierazi digute. Frantses *Wikipedia*n modu batera baino gehiagotara eman daitezke datu biografikoak, eta, mota jakin bateko patroiak inplementatu ditugunez, espero bezala, errendimendua behera doa. Doitasuna oso baxua da (0,18), informazioa galtzen delako. Doitasun gramatikala ere baxua da (0,37) batzuetan parafrasiak agertzen direlako jatorrizko esaldietan, eta horrek akats gramatikalak sortu ditu gure hurbilpenean. Edonola ere, estaldura ona da (0,80) eta abiapuntu egokia da hobekuntzak egite aldera. κ neurria oso baxua atera zaigu (0,39), Cohenen kappak instantzia gutxi daudenean desadostasunak penalizatzen dituelako.

Galegoaren kasua ezberdina da zeharo. *Wikipedia*ko gidalerroetan ez da esaten nola adierazi behar diren datu biografikoak eta, gainera, aurkitutako parentetikoak gutxi eta *Euskal Wikipedia*n aurkitzen direnetatik ezberdinak dira. Hala ere, Biografix probatu dugu, eta errendimenduko doitasuna oso baxua den arren (0,15), sortutako esaldiak nahiko zuzenak dira (0,62). Gure ustez, egokitzapen ona egiteko galiziar *Wikipedia* aztertu beharko litzateke eta horren arabera Biografix moldatu.

Biografixen bertsio alemana laginean aurkitutako esaldi guztiak sinplifikatzeko gai izan da, eta bere estaldura altua izan da (0,88). Puntu ahula errendimenduko doitasuna izan da (0,60), beste hizkuntzetan ere gertatu den bezala. Sortutako esaldiak ere onargarriak (0,71) dira. Kasu honetan, bi hizkuntzalariak ados egon dira galdera guztien erantzunetan. Oro har, egokitzapen ona izan dela iruditu zaigu.

Azkenik, gaztelaniako egokitzapenaren kasuan, informazio-galera handia dela-eta, doitasuna oso baxua (0,33) da. Hala ere, doitasun gramatikala onargarria (0,67) da. κ koefizientea beste hizkuntzetakoa baino altuagoa (0,52) den arren, adostasun-proportzioa ez dago euskarazko kasuan lortutakotik urrun (88,76). Azpimarratzekoa da gaztelania aspaldidanik normalizatutako eta erregulatutako hizkuntza izanda ere, soilik jatorrizko esaldien % 85,00 dela zuzena, eta, datu biografikoak aurkezteko beste formatu batzuk dauden arren, lagineko esaldi guztietan euskarazko formatua aurkitu dugula (landutakoak % 100,00).

Ebaluatzaileen arteko desadostasun handienak gramatika eta puntuaziomarkak ebaluatzeko irizpideetan aurkitu ditugu. Batzuentzat aditzik gabeko esaldiak zuzenak ziren elipsia zegoela kontsideratu dutelako, eta beste batzuentzat, ordea, ez. Gure helburua sinplifikatzea denez, pentsatzen dugu esaldi sinplifikatu guztiek aditz jokatu bat gutxienez izan behar dutela. Ebaluatzaileek ez dute arazo handirik izan *Prog* galderak erantzuteko, eta

horregatik gure ebaluaziorako metodologia ona dela iruditu zaigu. Gainera, errore-analisia azkartu eta errazten du. Azpimarratu nahi dugu gure kasuan κ ez dela neurri egokiena izan, baina horixe erabili dugu datuen fidagarritasuna neurtzeko koefiziente estandarra delako. Ezin dugu ahaztu, oraindik badagoela aukera Biografixen emaitzak eta egokitzapenak hobetzeko, eta emaitzak hain onak izan ez diren hizkuntzetan (galegoa, gaztelania eta frantsesa) analisi sakonagoa egin ondoren, Biografix hizkuntza berrietan informazio biografikoa duten egitura parentetikoak sinplifikatzeko abiapuntu egokia izan daitekeela.

Amaitzeko, esan nahi dugu euskararako diseinatutako erregelak beste hizkuntzetan ere erabilgarriak izan direla.

Biografixen ebaluazio estrintsekoa

Ebaluazio estrintsekoa egiteko Senekori (Lopez-Gazpio eta Maritxalar, 2013) pasatu dizkiogu jatorrizko esaldiak eta sinplifikatutako esaldiak. Mota bakoitzarekin sortutako galderen kopuruak 6.8 taulan ikus daitezke. Kopuru horiek kasu-marken arabera ere eman ditugu.

Sarrera	Guztira	Absolutiboa	Inesiboa	Genitiboa	Bestelakoak
Jatorrizko esaldiak	34	23	7	2	2
Sinplifikatutako esaldiak	142	65	66	8	3

6.8 taula – Senekok sortutako galdera-kopurua jatorrizko esaldiekin eta esaldi sinplifikatuekin

Jatorrizko esaldiak erabilita, Seneko 34 galdera sortzeko gai da; esaldi bakoitzeko galdera bat gutxi gorabehera. Horietatik 23 absolutibo kasuarentzat osatu dira (gogoan izan kasu honek subjektua eta predikatiboa hartzen dituela) eta 7 sortu dira inesiboarentzat. Kontuan izanda informazio biografikoa lantzen ari garela, emaitza txarra dugu inesiboak denbora- eta toki-erlazioak zehazten baititu; hau da, oso galdera gutxi sortu dira *non* eta *noiz* galdetzaileekin. Bestalde, sarrera bezala esaldi sinplifikatuak erabilita, 65 galdera sortu dira absolutiboarekin eta 66 inesiboarekin. Sortutako galderen adibideak 6.9 taulan ikus daitezke.

Beraz, Biografix erabilita Senekoren emaitza hobetzea lortu dugu, batez ere denbora- eta toki-galderak sortuz datu biografikoak dituzten testuetan. Honenbestez, Biografixen erabilgarritasuna HPko aplikazio aurreratu batean frogatu dugu.

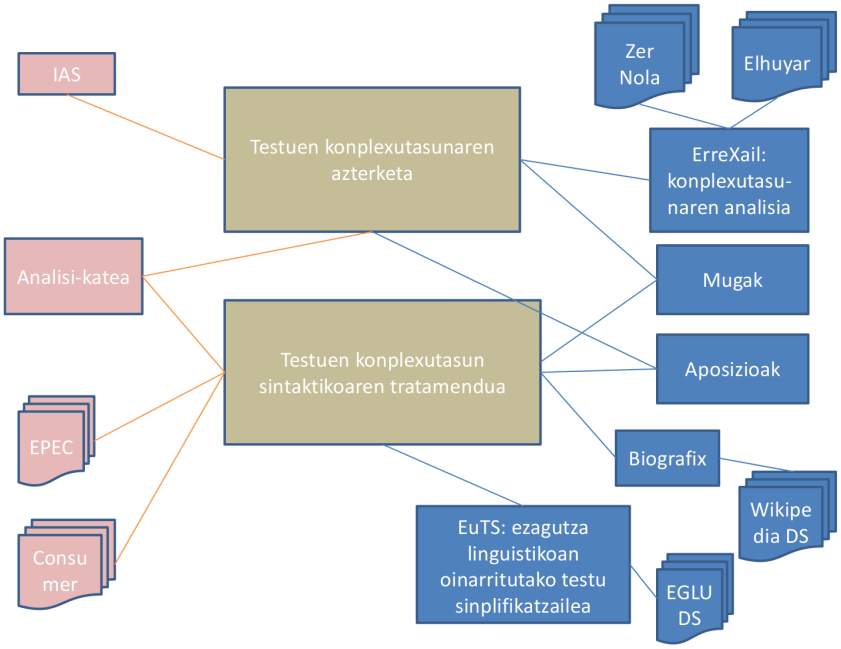
Jatorrizko esaldia	**Sinplifikatutako esaldia**
Eduardo Hughes Galeano (Montevideo, 1940ko irailaren 3a -) Uruguaiko kazetari eta idazlea da.	*Eduardo Hughes Galeano Uruguaiko kazetari eta idazlea da. Eduardo Galeano 1940ko irailaren 3an Montevideon jaio zen.*
Sortutako galdera	**Sortutako galdera**
Nor da Eduardo Hughes Galeano Montevideo 1940ko irailaren 3a?	*Non jaio zen Eduardo Galeano? Nor jaio zen 1940ko irailaren 3an Montevideon?*

6.9 taula – Senekok sortutako galderen adibideak

6.5 Laburpena

Kapitulu honetan EuTS sistemaren diseinua aurkeztu dugu. EuTS sistema euskarazko sinplifikazio automatikoa egingo duen sistema da, eta bi sinplifikazio-mota egingo ditu: ordezkapen sintaktikoen sinplifikazioa eta sinplifikazio sintaktikoa. Sistema hori osatuko duten moduluak azaldu ditugu hemen. Ordezkapen sintaktikoen sinplifikazioa egiten duen modulua inplementatu dugu, eta sinplifikazio sintaktikoa egingo duten moduluek osatuko dituzten tresnak informazio linguistikoarekin hornitu ditugu. Horretaz gain, EuTSren arkitektura eta diseinatutako erregelak probatzeko, egitura parentetikoen kasu-azterketa egin dugu. Egitura horiek sinplifikatzeko Biografix tresna sortu dugu. Erregelak inplementatu ditugu eta tresnaren ebaluazio intrintsekoa eta estrintsekoa egin ditugu. Patroietan oinarritutako tresna denez, beste hizkuntza batzuetara egokitu dugu. Horrela, euskarazko erregelak beste hizkuntzetara egoki daitezkeela ikusi dugu, inplementazioan lan gehiago egin behar den arren. Kapitulu honetan aurkeztutako sistemak eta datu-multzoak 6.5 irudian gehitu ditugu.

Hurrengo kapituluan (7. kapituluan), berriz, eskuz sinplifikatutako testuen analisia aurkeztuko dugu guk proposatzen ditugun hurbilpena eta eragiketak konfrontatzeko.



6.5 irudia – Tesian erabilitako baliabideak eta tresnak, eta egindako ekarpenak

7

Euskarazko Testu Sinplifikatuen Corpusa (ETSC)

Kapitulu honetan Euskarazko Testu Sinplifikatuen Corpusa (ETSC) aurkeztuko dugu. Corpusa nola osatu dugun eta nola etiketatu den deskribatuko dugu. Halaber, etiketatze horren emaitzak eta ondorioak emango ditugu eta ahal dugun guztietan euskarazko corpusa beste hizkuntzetako corpusekin alderatuko dugu.

7.1 Sarrera

Euskarazko Testu Sinplifikatuen Corpusa (ETSC) eskuz sinplifikatutako testuekin eta euren jatorrizko bertsioarekin osatu dugun testu-bilduma da. Mota horretako corpusa biltzearen helburua izan da, batetik, gure hurbilpena (3. eta 4. kapituluetan azaldutakoa) beste hurbilpen batzuekin kontrastatzea eta, bestetik, beste hurbilpen batzuk euren artean erkatzea. Konparazio horiek egiteko, etiketatze-eskema osatu dugu testuak sinplifikatzean egin diren eragiketak deskribatzeko. Testuak BRAT tresnaren (Stenetorp *et al.*, 2012) bitartez etiketatu ditugu.

Testu sinplifikatuen corpusek bi helburu nagusi izan dituzte erdal hizkuntzetako lanetan: alde batetik, jatorrizko testuetatik testu sinplifikatuak lortzeko zein eragiketa egin diren aztertzea, ondoren eragiketa horiek automatizatzeko, eta, beste aldetik, ikasketa automatikoko teknikak erabiltzen dituzten sistemak sortzeko baliabideak izatea. Esan bezala, euskarazko cor-

pusa sortzearen helburuak izan dira guk proposatu dugun hurbilpena beste batzuekin alderatzea (ebaluazioa), zein eragiketa egin diren aztertzea, beste hurbilpen batzuk euren artean erkatzea eta testuak sinplifikatzean amankomunak diren irizpideak aurkitzea. Etorkizunean ez dugu alde batera uzten ikasketa automatikoko teknikak erabiltzeko baliabidea izatea.

7.2 ETSC corpusaren osaera eta etiketatzea

ETSC osatzeko, *Elhuyar (T-comp)* corpusetik 30 erreportaje hartu ditugu. Erreportaje horiek aldez aurretik ErreXail sistema entrenatzeko erabili ditugu, testu konplexuen atalean, hain zuzen ere. Testu guztiak zientzia-dibulgazio domeinukoak dira eta 3 jakintza-arlo hauetan banatzen dira: Giza Zientziak eta Historia, Medikuntza eta Teknologia.

Jakintza-alor bakoitzeko 10 testu daude eta jatorrizko testu horiek dituzten esaldi- eta hitz-kopurua aurkezten dugu 7.1 taulan.

Jakintza-alorra	**Esaldi-kopurua**	**Hitz-kopurua**
Giza Zientziak eta Historia	257	12.307
Medikuntza	231	12.231
Teknologia	176	10.199

7.1 taula – Esaldi- eta hitz-kopurua jatorrizko testuetan

Jatorrizko 30 testu horiek testuak sinplifikatzeko atazan egokiak direla probatzeko, jakintza-alor bakoitzetik testu bat hartu dugu eta eskuz sinplifikatu dugu. Sinplifikazio-lan horretatik ateratako ondorioekin orientaziorako gidalerroak idatzi ditugu. Fase honi *abiapuntua* edo *proba-fasea* deitu diogu.

Ondoren, alor bakoitzeko beste testu bat bi sinplifikatzaileri[1] eman dizkiegu euren arteko berdintasunak eta ezberdintasunak ikusteko asmoz. Bi sinplifikatzaile horiei dagokienez, bata euskara irakaslea izan zen euskaltegian eta testuak ikasleen mailara egokitzen zituen (irakaslea), eta, bestea lan hau egin arte sinplifikazioarekin harremanik izan ez duen itzultzailea da (itzultzailea).

Fase horri *konfrontazioa* edo *fintze-fasea* deitu diogu. Azkenik, konfrontazio-fasean ateratako ondorioekin, bete behar diren eragiketen zerrenda osatu

[1]Kapitulu honetan, sinplifikatzaileak, testuak eskuz sinplifikatu dituzten pertsonak dira.

dugu corpusaren eskuzko sinplifikazio sistematikoa egiteko. Fase horri *finkatzea* edo *handitze-fasea* deitu diogu, eta zerrenda horri jarraituz sinplifikatuko ditugu gainontzeko testuak. Jarraian, fase horretako bakoitzean zer egin den zehatzago adieraziko dugu:

1. **Abiapuntua (proba-fasea).** *Elhuyar (T-comp)* corpuseko testuak gure atazarako baliagarriak ote diren ikusteko, hiru testu eskuz sinplifikatu ditugu (alor bakoitzetik, bat). Fase honetan, batez ere kontuan izan ditugu, batetik, beste hizkuntzetan proposatu diren irizpideak eta, bestetik, sinplifikazioa automatikoki egiteko aurreko kapituluetan proposatu ditugun erregelak, azterketa linguistikoan oinarritutakoak (7.2 taula, iturria guk).

   Beste hizkuntzetatik hartu digun irizpideak irakurketa errazean oinarritzen dira. Irakurketa errazeko irizpide horiek oso orokorrak dira (Bott eta Saggion, 2014) batez ere sistema automatikoak eraikitzeari begira. Hala ere, lan askoren abiapuntua dira eta sinplifikazioa gauzatuko duten sinplifikatzaileei lana errazteko asmoz jaso ditugu 7.2 taulan (iturria, (Mitkov eta Štajner, 2014). Esaterako, Mitkov eta Štajner-ek (2014) irakurketa errazeko testuak ingelesez idazteko, hiru gidalerroen konparazioa, gaztelaniako irizpideak eta irizpide propioak aurkezten dituzte.

Irizpidea	**Iturria**
Erabili esaldi sinpleak eta laburrak	(Mitkov eta Štajner, 2014)
Kendu anaforak	(Mitkov eta Štajner, 2014)
Erabili soilik maiztasun handiko hitzak	(Mitkov eta Štajner, 2014)
Erabili beti hitz bera gauza bat adierazteko	(Mitkov eta Štajner, 2014)
Mantendu hurrenkera kanonikoa eta logikoa	Guk
Berreskuratu subjektu eta objektu elidituak, beharrezkoa bada	Guk
Berreskuratu aditz elidituak	Guk
Ideia edo gai bakarra esaldi bakoitzean, aditz jokatu bakarra esaldi bakoitzean	Guk

7.2 taula – Irakurketa errazeko irizpideak eta guk gehitutakoak

   Proba-fase honetan eragiketen zerrenda (7.3 taula) ere osatu dugu. Eragiketa horiek, oro har, TSAko lanetan erabili izan dira, baina azkeneko

zazpiak guk gehitu ditugu, fase honetako testuak sinplifikatzean aurkitu diren eragiketak baitira. Eragiketa horiekin corpusaren etiketatze-eskemaren lehenengo bertsioa (protoeskema; CBTS-anotationScheme-v0) osatu dugu.

Eragiketa	**Iturria**
Perpaus koordinatuak banatu	TSA
Mendeko perpausetatik esaldi berriak sortu	TSA
Aposizio-sintagmetatik esaldi berriak sortu	TSA
Egitura parentetikoak ezabatu, informazioa kenduz	TSA
Egitura parentetikoetatik esaldi berriak sortu	TSA
Izenordainak ordezkatu	TSA
Hitzen/sintagmen hurrenkera aldatu	TSA
Hitzak eta sintagmak ordezkatu (sinonimo bat, hitz ezagunago bat)	TSA
Ahots pasiboak aktibo bihurtu	TSA
Perpausen hurrenkera aldatu	TSA
Lokailuak aldatu	TSA
Parafrasiak egin	TSA
Eragiketarik ez	TSA
Subjektuak eta objektuak esplizitu egin	Guk
Aditz eliditua esplizitu egin	Guk
Estilo zuzenera pasatu	Guk
Puntuazio-markak aldatu	Guk
Pertsona aldatu	Guk
Bestelakoak	Guk

7.3 taula – Protoeskeman (CBTS-anotationScheme-v0) jasotako eragiketak

Proba-fasean jatorrizko testuak gure atazarako baliagarriak direla frogatzearekin batera, testuak sinplifikatzeko metodologia ezarri dugu testuak sinplifikatu behar dituztenak jarrai dezaten:

- Irakurketa errazeko irizpideak eta gureak irakurri
- Sinplifikatu beharreko testua irakurri

- Testuan egon daitezkeen oztopoak edo egitura konplexuak markatu (aukerakoa)
- Testua esaldiz esaldi sinplifikatu (beharrezkoa bada) eta egin den eragiketa adierazi (aukerakoa)
- Sinplifikatutako testua berrirakurri eta errepasatu

2. **Konfrontazioa (fintze-fasea).** Konfrontazioa egiteko bi sinplifikatzaileei hiru testu eman dizkiegu, goian aipatutako eragiketekin eta irizpideekin batera. Fintze-fasearen helburua da, alde batetik, ikuspuntu edo hurbilpen ezberdina duten bi pertsona zenbateraino dauden ados ikustea (baldin badaude), eta, beste aldetik, guk emandako eragiketez gain zein beste eragiketa egiten dituzten ikustea gure hasierako etiketatze-eskema edo protoeskema[2] osatzeko. Fasearen amaieran, testuen jatorrizko bertsioaz gain, irakasleak eta itzultzaileak sinplifikatutako testu bakoitzaren bertsioak izango ditugu. Adibide gisa, sinplifikatutako testu horietako baten ("Etxeko") testuko lehen paragrafoa hiru bertsiotan ematen dugu 7.4 taulan.

Jatorrizko testua	**Itzultzaileak sinplifikatutakoa**	**Irakasleak sinplifikatutakoa**
Minbizi gehienen jatorria edo eragilea ezezaguna bada ere, jakina da kasu batzuek jatorri genetikoa dutela.	Minbizi gehienen jatorria edo eragilea ezezaguna da. Kasu batzuek jatorri genetikoa dute.	Zergatik edo nola sortzen da minbizia? Ez dakigu orokorrean. Kasu batzuetan heredatu egiten da.
Kasu horiek identifikatuz gero, gaitzaren garapena eragozteko neurriak hartzeko aukera dago.	Kasu horiek identifikatzen dira; orduan, gaitzaren garapena eragozteko neurriak har daitezke.	Kasu horiek identifikatzen badira, neurriak hartzea badago minbiziaren gaitzak aurrera ez egiteko.
Horretan dabiltza, beste batzuen artean, Donostiako Onkologikoko Genetika Saileko sendagileak.	Identifikazio lanetan dabiltza, beste batzuen artean, Donostiako Onkologikoko Genetika Saileko sendagileak.	Donostiako Onkologikoko sendagileak neurri horiek aztertzen ari dira.

(Jarraipena hurrengo orrialdean)

[2] Mitkov eta Štajner-en (2014) lanetan ematen dituzten eragiketak oso lausoak eta nahikoak ez direla iruditu zaigunez, gure helburua ahalik eta eragiketa posible gehienak lortzea izan da.

Jatorrizko testua	**Itzultzaileak sinplifikatutakoa**	**Irakasleak sinplifikatutakoa**
Aldi berean, ikertzaileek minbizi horiek hobeto ezagutzeko ahaleginetan dabiltza.	Aldi berean, ikertzaileak minbizi horiek hobeto ezagutzeko ahaleginetan dabiltza.	Aldi berean, ikertzaileak ere minbizi horiek hobeto ezagutzeko lanetan ari dira.

7.4 taula – Paragrafo baten konparazioa, esaldika lerrokatuta

Behin testu sinplifikatuak ditugula:

2.1. Corpusaren hasierako etiketatzea eta analisia egin dugu. Etiketatze hori paragrafo-mailan egin dugu eta protoeskemako eragiketetan oinarritu gara. Testuetan bestelako eragiketak aurkitu ditugunez, eragiketa horiek aztertu ditugu, eta eragiketetan sailkatu eta azpisailkapenak egin ditugu.

2.2. Protoeskema (CBTS-anotationScheme-v0) osatu dugu eta behin-behineko etiketatze-eskema (CBTS-anotationScheme-v1) sortu dugu. Behin-behineko etiketatze-eskema eragiketaren mailaren arabera antolatu dugu; maila horiek dira: sintaxia, lexikoa, hurrenkera-aldaketa, diskurtsoa, elipsiaren tratamendua, informazioaren tratamendua, eta bestelakoak. Horretaz gain, eragiketa bakoitza zein motatakoa den adierazi dugu; mota horiek dira: banaketa, transformazioa, hurrenkera-aldaketa, ezabatzea, txertatzea, bestelakoak eta eragiketarik eza. Eskema hori 7.5 taulan ikus daiteke.

2.3. Bi sinplifikatzaileen arteko lehen erkaketa ere fase honetan egin dugu.

Eragiketaren maila	**Mota**
Eragiketa sintaktikoak	
Koordinazioak banatu eta esaldi berriak sortu	banaketa/transformazioa
Mendeko perpausak banatu eta esaldi berriak sortu	banaketa/transformazioa
Aposizio-sintagmak banatu eta esaldi berriak sortu	banaketa/transformazioa
Egitura parentetikoak banatu eta esaldi berriak sortu	banaketa/transformazioa
Ahots pasiboa edo inpertsonala ahots aktibo bihurtu	transformazioa
Mendeko perpausa perpaus koordinatu bihurtu	transformazioa
Aditz ez-jokatuak jokatu bihurtu	transformazioa

(Jarraipena hurrengo orrialdean)

Eragiketaren maila	**Mota**
Aditz jokatuak ez-jokatu bihurtu	transformazioa
Eragiketa lexikalak	
Hitzak edo sintagmak sinonimo batekin edo hitz ezagunago batekin ordezkatu	transformazioa
Hurrenkera aldatzeko eragiketak	
Hitzen/sintagmen hurrenkera aldatu	hurrenkera-aldaketa
Perpausen hurrenkera aldatu	hurrenkera-aldaketa
Diskurtsoko eragiketak	
Izenordain-mailako korreferentzia ebatzi	transformazioa
Diskurtso-markatzailea aldatu	transformazioa
Diskurtso-markatzailea ezabatu	ezabatzea
Diskurtso-markatzailea txertatu	txertatzea
Diskurtso-markatzaile grafikoak txertatu	txertatzea
Zehar-estiloa estilo zuzen bihurtu	transformazioa
Diskurtsoaren pertsona aldatu	transformazioa
Elipsiaren tratamendua	
Elidítutako subjektuak eta objektuak berreskuratu	txertatzea
Elidítutako aditzak berreskuratu	txertatzea
Informazioaren tratamendua	
Egitura parentetikoak ezabatu	ezabatzea
Alde batera utzitako informazioa berreskuratu	txertatzea
Esaldiak trinkotu	bateratzea
Erredundantziak ezabatu	ezabatzea
Bestelakoak	
Parafrasiak egin	transformazioa
Puntuazio-markak aldatu	transformazioa
Zuzenketak egin	transformazioa
Bestelakoak	bestelakoak
Eragiketarik ez	eragiketarik ez

7.5 taula – Behin-behineko etiketatze-eskema (CBTS-anotationScheme-v1)

2.4. Eragiketa jakinen analisi sakona egin dugu, behin-behineko etiketatze-eskema horretan eta italierako corpuseko eskeman (Brunato

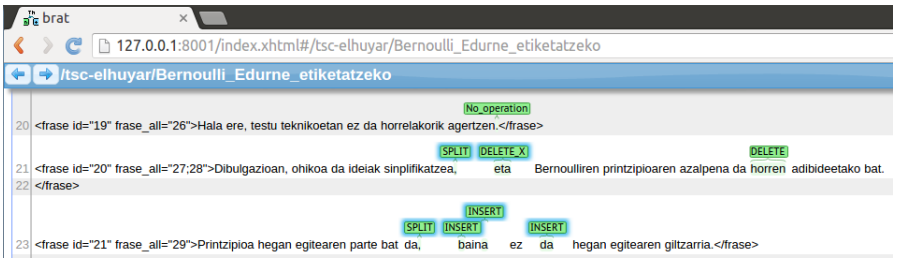
et al., 2015) izen berarekin edo izen ezberdinarekin deitzen ziren fenomenoak bat datozen ikusteko.

2.5. Etiketatze-eskemak mapeatu eta osatu ondoren, behin betiko etiketatze-eskema definitu dugu. Etiketatze-eskema hori 7.3 azpiatalean azalduko dugu.

2.6. Testuak esaldiz esaldi lerrokatu ditugu.

2.7. Testuak BRAT tresnarekin etiketatu ditugu etiketatze-eskemaren arabera .

Etiketatutako testu zati bat 7.1 irudian ikus daiteke. Irudi horretan "Bernoulli" izeneko jatorrizko testua esaldika banatuta ikusten da eta testuaren gainean testu sinplifikatua sortzeko egindako eragiketaren etiketa ageri da.



7.1 irudia – BRATen etiketatutako testu baten zatia

Behin testuak etiketatu ditugunean:

2.8. Eragiketen intzidentzien emaitzak atera ditugu.

2.9. Hurbilpenen konfrontazioa egin dugu.

2.10. Ondorioak atera ditugu.

Etiketatzearen emaitzak 7.4 atalean azalduko ditugu.

3. **Finkatzea (handitze-fasea).** Itzultzailearen eta irakaslearen testuak analizatu ondoren, biek komun dituzten eragiketekin testuak eskuz sinplifikatzean bete beharreko eragiketen zerrenda osatu dugu (C eranskina). Zerrendan oinarrituz, corpusean bildu ditugun gainontzeko testuak sinplifikatuko ditugu. Etorkizunean ere beste domeinuetako testuak gehitzea pentsatu dugu, albisteak edo *Wikipedia*ko testuak, esaterako.

7.3 Etiketatze-eskema

Etiketatze-eskemen bitartez jatorrizko testuak sinplifikatzean egindako eragiketak sistematikoki antolatzen dira (Bott eta Saggion, 2014; Brunato *et al.*, 2015). Eragiketa horiek sinplifikazioan emandako aldaketak deskribatzeko eta analizatzeko balio dute. Gure etiketatze-eskeman eragiketak makroeragiketetan bildu ditugu (7.13 taula). Makroeragiketa horiek TSAko lan gehienetan eta testu sinplifikatuen corpus gehienetan aurki daitezke. Ondorengo puntuetan makroeragiketa bakoitza aurkezten dugu euskararako egokitu dugun definizioa ez ezik, gehitu ditugun beste bi makroeragiketa zein diren ere adieraziz: *no_operation* eta *other*.

- Ezabatzea (*delete*): kasu-markak, hitzak, sintagmak, perpausak edo esaldiak ezabatzea (menderagailuak ez ditugu kontuan hartzen, transformazioetatik atera baitaitezke)
- Bateratzea edo fusioa (*merge*): perpaus/esaldi bat baino gehiagotik perpaus/esaldi bat sortzea
- Banaketa (*split*): sintagmak, perpausak edo esaldiak banatzea
- Transformazioa (*transformation*): jatorrizko hitzak, sintagmak, perpausak edo esaldiak eraldatzea
- Txertaketa (*insert*): elementu berriak (hitzak, sintagmak, perpausak edo esaldiak) txertatzea
- Hurrenkera-aldaketa (*reordering*): hitzen, sintagmen, perpausen edo esaldien hurrenkera aldatzea
- Eragiketarik eza (*no_operation*): eragiketarik ez egitea
- Bestelakoak (*other*): bestelakoak (etiketatze-eskemak jasotzen ez dituenak), edo eragiketen nahasketa (bat baino gehiago aldi berean gertatzen direnak edo sailkatzeko zailak direnak)

Hurrengo azpiataletan zehatzago ikusiko dugu makroeragiketa bakoitzean zer hartzen dugun kontuan eta zein eragiketa biltzen ditugun. Corpusetik ateratako adibideak ere jarriko ditugu. Adibide horiek tauletan emango ditugu eta unean nabarmendu nahi dugun fenomenoa letra etzanez adieraziko

dugu. Kontuan izan behar da zaila dela eragiketa bakarra duen adibidea aurkitzea eta, horregatik, adibide batzuetan eragiketa bat baino gehiago agertuko dira.

7.3.1 Ezabatzea (*delete*)

Ezabatze edo *delete* makroeragiketarekin jatorrizko testutik elementuak ezabatzea adierazi nahi da. Bi motako ezabatzeak bereizten ditugu ematen duten informazioaren arabera:

- Informazioaren ezabatzea (*delete_ info*): informazioa edo unitate lexikalak ezabatzea. Informazioaren ezabatzeak informazioaren tratamenduarekin lotuta daude, hau da, ezabatze horietan informazioa galtzen da. Eragiketa hori, gure ustez, laburpen automatikoetatik gertuago dago eta ez dator bat gure sinplifikazioaren hurbilpenarekin. Hala ere, corpusean eragiketa hori jasotzea beharrezkoa zaigu sinplifikatzaileek egindako eragiketa delako eta beste hizkuntzetan ere azaltzen delako. 7.6 taulako adibidean informazioa duen (agian ez garrantzitsua) *Sortzen den* erlatibozko perpausa ezabatu da.

- Ezabatze funtzionala (*delete_ fuctional*): hitz edo token funtzionalen ezabatzea, hau da, juntagailuak, indartzaileak, lokailuak, kasu-markak eta puntuazio-markak ezabatzea. Ezabatze funtzionalek ez dute eraginik oro har informazioaren osotasuna mantentzean, ñabardurak ken ditzaketen arren. 7.6 taulako adibidean ikus daitekeen bezala, *Eta* juntagailuaren ezabatzearekin ez da informaziorik galtzen.

Eragiketa	**Jatorrizkoa**	**Sinplifikatutakoa**
Informazioa ezabatzea	*Sortzen den* aldea oso handia da	Aldea oso handia da
Ezabatze funtzionala	*Eta* beste edozein hegazkinekin ere gauza bera gertatzen da	Beste edozein hegazkinekin ere gauza bera gertatzen da

7.6 taula – Ezabatze-makroeragiketaren adibideak

7.3.2 Bateratzea (*merge*)

Bateratze, fusio edo *merge* makroeragiketarekin, elementuen bateratzea edo fusioa adierazi nahi da. Makroeragiketa hori ez dugu corpusean maiz aurkitu, eta, ondorioz, ezin izan dugu eragiketarik edo azpieragiketarik definitu. Azpisailkapena egite aldera, esaldien bateratzea, sintagmen bateratzea eta hainbat mailatako bateratzeak bereiziko genituzke, baina ez dugu ebidentzia nahikorik oraingoz sailkapena sakontzeko. 7.7 taulan bateratutako bi esaldien adibideak ikus ditzakegu eta bateraketa hori egiteko *Haien* izenordainaren oinarria erabili da.

Eragiketa	**Jatorrizkoa**	**Sinplifikatutakoa**
Bateratzea	Adibide bat gaur egungo hegazkin komertzialen hegoak dira. *Haien* diseinua plano aerodinamiko superkritikoan oinarrituta dago.	Gaur egungo hegazkin komertzialen hegoen diseinua plano aerodinamiko superkritikoan oinarritzen da.

7.7 taula – Bateratze-eragiketaren adibidea

7.3.3 Banaketa (*split*)

Banaketa edo *split* makroeragiketa perpausen, sintagmen edo morfemen banaketa egitea da hainbat perpaus edo esaldi sortzeko helburuarekin. Hau da, egitura bat banatzen denean eta hortik aditza duen beste egitura bat sortzen denean gertatutako banaketa da.

Banaketa ezberdinak bereizten ditugu indarra eta fenomenoaren arabera:

1. Indarra: banaketaren indarra banaketa egiteko erabili den puntuazio-markaren araberakoa da. Banaketa leuna edo *split-soft* banaketa erabiltzen da puntu eta koma edo koma erabili bada. Puntuarekin egin bada, aldiz, banaketa gogortzat (edo arruntzat) edo *split-hard* banaketatzat jotzen da. Banaketa indarren arabera sailkatzea erabaki dugu, ikusi baita banaketa leun horiek banaketaren definizioarekin bat datozela eta ez beste fenomenoren batekin.

2. Fenomenoa: banatzen den fenomenoa zehazteko sailkapen morfosintaktikoak erabili ditugu. Koordinazioa, perpaus osagarriak, erlatibozko perpausak, perpaus adberbialak, aposizio-sintagmak, postposizioak eta bestelakoak bereizi ditugu.

Banaketak etiketatzean, bi irizpideak hartzen dira kontuan. Banaketen adibideak 7.8 taulan ikus ditzakegu.

Eragiketa	**Jatorrizkoa**	**Sinplifikatutakoa**
Gogorra-perpaus koordinatuak	Dibulgazioan, ohikoa da ideiak sinplifikatzea, eta Bernoulliren printzipioaren azalpena da horren adibideetako bat.	Dibulgazioan ohikoa da ideiak sinplifikatzea. Bernoulliren printzipioaren azalpena da adibideetako bat.
Leuna-perpaus koordinatuak	Hortik aurrerako azalpena konplexua da, eta hegalari batetik bestera asko aldatzen da.	Hortik aurrerako azalpena konplexua da*;* hegalari batetik bestera asko aldatzen da.

7.8 taula – Indarraren araberako banaketen adibideak

7.3.4 Transformazioa (*transformation*)

Transformazio makroeragiketarekin, jatorrizko hitzaren edo egituraren eraldaketa jaso da. Mota ezberdinetakoak daude: transformazio lexikalak, morfologikoak, sintaktikoak, diskurtsokoak, zuzenketak eta bestelakoak. Etiketatze-eskemako eragiketa gehienak mota honetakoak dira eta motaren arabera sailkatu ditugu.

- **Lexikalak:** Subst_Syn (sinonimoen ordezkapena) eta Subst_Multi-Word (sintagmen ordezkapena)
- **Morfologikoak:** Pas2Act (pasiboa -> aktiboa edo inpertsonala-> pertsonala), Fin2NonFin (aditz jokatua -> ez-jokatua), NonFin2Fin (aditz ez-jokatua -> jokatua), Subst_Per (pertsona aldatu) eta Verb_Feats (aldaketak aditzean)
- **Sintaktikoak:** Clause2Phrase (perpausa -> sintagma), Phrase2Clause (Sintagma-> perpausa), Ind2Dir_Speech (estilo aldaketa: zehar -> zuzen), Dir2Ind_Speech (estilo aldaketa: zuzen -> zehar), Sub2Main (mendekoa-> nagusia), Main2Sub (nagusia-> mendekoa), Connect_Syntax (sintaxi konektorea aldatu) eta Sub2Coor (mendekoa koordinatu bihurtu)
- **Diskurtsokoak:** Coref (korreferentziaren ebazpena) eta Connect_Disc (lokailua aldatu)

- **Zuzenketak:** Correction (akats ortografikoen eta gramatikalen zuzenketak)
- **Hainbat:** Reform (erreformulazioa edo parafrasiak) eta Other_Subst (bestelakoak, zehatu beharrekoak)

Corpusean aurkitu ditugun transformazioen adibideak 7.9 taulan jaso ditugu. Baliteke taulan azaldu nahi dugun eragiketaz gain, adibideren batean beste eragiketaren bat agertzea; izan ere, ez da erraza kasu batzuetan eragiketa bakarra aurkitzea.

Eragiketa	**Jatorrizkoa**	**Sinplifikatutakoa**
Subst_Syn	ahaleginetan	lanetan
Subst_MultiWord	urteetan zehar	urtero
Pas2Act	ikusi da	ikusi dute
Fin2NonFin	hegazkin horiei airean eusten dien printzipio fisikoa	hegazkin horiei airean eusteko printzipio fisikoa
NonFin2Fin	Airea beherantz bultzatuta	Airea beherantz bultzatzen da
Subst_Per	orduan odolean begiratzen dugu	orduan odolean begiratzen dute
Verb_Feats	gai izango litzateke	gai izango da
Clause2Phrase	Jatorri genetikoa duten minbizi gehienetan	Jatorri genetikodun minbizi gehienetan
Phrase2Clause	bakoitzak oso diseinu ezberdinarekin	Bakoitzak bere diseinua du
Ind2Dir_Speech	familian zenbat kasu dauden galdetzen dugu	zenbat kasu daude familian?
Dir2Ind_Speech	horiekin “ez da eragozten” minbizia sortzea	horiekin ez dela galarazten minbizia sortzea
Sub2Main	fluxu horrek presio handiagoa egiten diola hegoari behetik goitik baino.	fluxu horrek presio handiagoa egiten dio hegoari behetik goitik baino.
Main2Sub	Familia barruan minbizi horietako kasu asko dituzten pertsonak iristen dira kontsultara.	Mujikak esan du kontsultara etortzen direla familia bereko pertsonak.
Connect_Syntax	angelu horren inguruan irauten duen bitartean	angelu horren inguruan irauten badu

(Jarraipena hurrengo orrialdean)

Eragiketa	**Jatorrizkoa**	**Sinplifikatutakoa**
Sub2Coor	Hartara, mutazioa identifikatuta,	Hartara, mutazioa identifikatzen dugu;
Coref	Mende hartan	XVIII. mendean
Connect_Disc	beraz	ondorioz
Correction	abiadura (...) izan beharko luke	abiadurak (...) izan beharko luke
Reform	Zama guztiarekin, 573 tonara irits daiteke.	Zama guztiarekin, 573 tona pisatzen du, gutxi gorabehera.
Subst_Other	hegaldiaren azalpenetik	hegaldiaren azalpenean

7.9 taula – Transformazio-eragiketen adibideak

7.3.5 Txertaketa (*insert*)

Txertaketa edo *insert* makroeragiketak testuan elementu berriak txertatzean gertatzen dira. Elementu berri horiek aldez aurretik ezabatutako erlazio bat berreskuratzeko (esaldi berrietan) edo elipsia tratatzeko (elidi tutakoak eta beharrezkoak ez direnak) txerta daitezke. Beraz, txertaketak bereizteko bi irizpide hartu ditugu:

1. Txertaketa egin den tokia: jatorrizko esaldia zena edo sinplifikatutako esaldia (esaldi berria)

2. Elipsi-mota: morfologikoki markatuta edo markatu gabe

Bi irizpide horiek kontuan hartuta hiru motatako eragiketa bereizi ditugu:

- Esaldi berrietan (*funct_NS*): esaldi berriak sortzean txertatu diren aditzak, argumentuak... Eragiketa hau ezin da gertatu aurretik jatorrizkoa zen esaldian banaketa-eragiketa egin ez bada.

- Elidi tutakoak (*ellided*): sintaxiak edo morfosintaxiak ahalbidetuta elidi tutako izenak edo aditzak, hau da, marka morfologikoa dutenak

- Beharrezkoak ez direnak (*non_required*): esanahia argiago egiteko berreskuratutako adjektiboak, adberbioak, esaldiak, elidi tutako objektu eta subjektuak...

Txertaketen adibideak 7.10 taulan ikus ditzakegu.

Eragiketa	Jatorrizkoa	Sinplifikatutakoa
Esaldi berrietan	La Pizarra de Yuri blogeko Antonio Cantó dibulgatzaile eta hegazkinetan adituak	*Antonio Cantó* dibulgatzailea *da*; *Antonio Cantó* hegazkinetan aditua *da*. La Pizarra de Yuri blogeko blogaria da.
Elidituta-koak	endometrioko minbiziaren pronostikoa obulutegietakoa baino askoz ere hobea izaten da	endometrioko minbiziaren pronostikoa *obulutegietako minbiziaren pronostikoa* baino askoz ere hobea izaten da
Beharrezkoak ez direnak	Eraldatuta badaude	*Proteinak* eraldatuta badaude

7.10 taula – Txerkaketa-eragiketen adibideak

7.3.6 Hurrenkera-aldaketa (*reordering*)

Hurrenkera-aldaketak bi irizpideren arabera antolatu ditugu:

1. Mugitu den elementua: sintagma, perpausa edo aditz laguntzailea
2. Tokia: jatorrizkoa zen esaldian edo jatorrizkoa zen esalditik sinplifikatutako esaldira edo (esaldi berrietarako mugimenduak).

Horiek kontuan izanik lau motatako hurrenkera-aldaketa eragiketa bereizi ditugu:

- Sintagmen hurrenkera-aldaketa (*Reord_Phrase*): tokiz aldatu diren sintagmak, baina perpaus berean mantendu direnak, hau da, jatorrizkoa zen esaldian
- Perpausen hurrenkera-aldaketa (*Reord_Clause*): tokiz aldatu diren perpausak, baina jatorrizkoa zen esaldian mantendu direnak
- Aditz laguntzaileen hurrenkera-aldaketa (*Reord_Aux*): aditz laguntzailea jatorrizkoa zen esaldian mugitzen denenan
- Sintagmak esaldi berrietan (*Reord_NS_Phrases*): esaldi berriak sortu direlako tokiz mugitu diren sintagmak. Mugimendu hori egiteko aurretik banaketa-eragiketa bat egin behar izan da.

Hurrenkera-aldaketa horien adibideak 7.11 taulan ikus daitezke.

Eragiketa	Jatorrizkoa	Sinplifikatutakoa
Sintagmak	(...) argitu du *Bachiller astronomoak*	*Bachiller astronomoak* argitu du
Perpausak	Aireak hegazkinaren inguruan duen jokabidea zoruak alda dezake, *hegaldia oso baxua denean.*	*Hegaldia oso baxua denean* zoruak hegazkinaren inguruko airearen jokabidea alda dezake.
Aditz laguntzaileak	Orain dela 25 urte, berriz, eguzki-sistemako planetak baino ez *ziren* ezagutzen.	Orain dela 25 urte, berriz, eguzki-sistemako planetak bakarrik ezagutzen *ziren.*
Sintagmak esaldi berrietan	Hala ere, adituek *aurreikusten dute* planetagaien % 90, gutxi gorabehera, benetako planetak izango direla.	Hala ere, planetagaien % 90, gutxi gorabehera, benetako planetak izango dira; *hala* aurreikusi dute adituek.

7.11 taula – Hurrenkera-aldaketako eragiketen adibideak

7.3.7 Eragiketarik eza (*no_operation*)

Esaldi batean eragiketarik egin ez denean, hau da, testu sinplifikatuan esaldia jatorrizkoan bezala mantendu den kasuetan erabili dugu makroeragiketa hau. Honela etiketatutako esaldiak interesgarriak izango dira eta etorkizuneko lanetan analizatu beharko ditugu jakiteko zergatik ez diren sinplifikatu, esaterako, perpaus bakunak direlako. Horrelako adibide bat 7.12 taulan eman dugu.

Eragiketa	Jatorrizkoa	Sinplifikatutakoa
Eragiketarik ez	Azken batean, hori da hegan egitearen sekretua.	Azken batean, hori da hegan egitearen sekretua.

7.12 taula – Eragiketarik ez eragiketaren adibidea

7.3.8 Bestelakoak (*other*)

Makroeragiketa honetan sartu ditugu orain arte etiketatze-eskeman jaso ez diren eragiketak, bestelako eragiketak, sailkatzeko zailak direnak. Soilik behar-beharrezkoa bada erabiltzekoa da eta honela etiketatutakoen analisia etorkizuneko lanetan egin beharko dugu, irtenbide berri bat emateko eta beharrezkoa bada, eragiketa berri bat sortzeko.

7.3 atalean aurkeztu dugun etiketatze-eskemaren laburpen eskematikoa jaso dugu 7.13 taulan.

Makroeragiketa	**Irizpidea**	**Eragiketak**
Ezabatzea	Informazioa	Informazioa vs. Hitz funtzionalak
Bateratzea		
Banaketa	Indarra	Gogorra vs. Leuna
	Fenomenoak	Perpaus koordinatuak
		Perpaus adberbialak
		Erlatibozko perpausak
		Aposizio-sintagmak/ Egitura parentetikoak
		Perpaus osagarriak
		Postposizioak
		Bestelakoak
Transformazioa	Mota	Lexikalak
		Morfologikoak
		Sintaktikoak
		Diskurtsokoak
		Zuzenketak
		Bestelakoak
Txertaketa	Tokia	Esaldi berrietan vs. jatorrizkoak zirenetan
	Elipsia	Elidi tutakoak markatuta vs. ez-beharrezkoak
Hurrenkera-	Tokia	Esaldi berrietan vs. jatorrizkoak zirenetan
aldaketa	Elementua	Sintagmak
		Perpausak
		Aditz laguntzaileak
Eragiketarik ez		
Bestelakoak		

7.13 taula – Etiketatze-eskemaren irudikapena

Beste hizkuntzetako etiketatze-eskemei dagokienez, gure etiketatze-

eskema italierako (Brunato *et al.*, 2015) eta gaztelaniako (Bott eta Saggion, 2014) etiketatze-eskemekin kontrastatu dugu (Brunato *et al.*, 2015) eta badira hainbat aldaketa. Lehenengo aldaketa erabilitako terminologian aurkitu dugu; 7.14 taulan jaso ditugu terminoen baliokidetzak.

Euskara	**Italiera**	**Gaztelania**
Makroeragiketak (*Macro-operations*)	Klaseak (*Classes*)	Lehenengo dimentsioa (*First dimension*)
Eragiketak (*Operations*)	Azpi-klaseak (*Sub-classes*)	Bigarren dimentsioa (*Second dimension*)

7.14 taula – Etiketatze-eskemetan erabilitako terminologia

Italierako etiketatze-eskemarekin alderatuz, makroeragiketa-mailan (beraiek *classes* deitzen dutena) guk *other* eta *no_operation* gehitu ditugu. Eragiketa-mailan (beraiek *sub-classes* deitzen dutena) ere ezberdintasunak daude. Adibidez:

1. "Ezabatzeetan" ezabatutako elementuaren kategoria zehazten dute; guk, ordea, horiek informazioaren tratamenduaren arabera sailkatzen ditugu (funtzionalak zein motatakoak diren ere etiketatzen dugun arren).

2. Txertaketetan zein elementu txertatu den begiratzen dute; guk hori ere egiten dugun arren, txertaketa-mota ezberdinak bereizten ditugu.

3. Transformazioetan, mailetako sailkapena ere egiten dute, baina espero den bezala, gure corpusean beraienean agertu ez diren eragiketa batzuk daude (eta alderantziz). Gurean aurkitu diren eragiketetako batzuk Clause2Phrase, Phrase2Clause, Main2Sub, Sub2Main, Sub2Coor eta Correction dira.

Gaztelaniako corpusa 145 esaldiko lagin esanguratsuaren gainean osatu dute eta bi dimentsioko taxonomia da. Lehenengo dimentsioan (gure makroeragiketaren parekoa) klase nagusiak agertzen dira eta bigarren dimentsioan (gure eragiketaren parekoa) klase horien zehaztapenak adierazten dira. Lehenengo dimentsioa gurearekin alderatuz, guk ez dauzkagun bi klase dituzte: *proximization* (informazioa irakurleari gerturatzea) eta *select* (informazioa nabarmentzea), eta gure *no_operation* eta *other* ez dira agertzen. Beste klase batzuen kasuan, erabilitako terminoa ezberdina da, baina eragiketa bera da. Adibidez, gure *transformation*i *change* deitzen diote eta *merge*ri *join*.

Bigarren dimentsioko motak ez dituzte zehazki azaltzen, baina ematen dituzten emaitzengatik ondoriozta dezakegu guk etiketatzen ditugun fenomenoen eta moten antzekoak direla. Adibidez, *change:lexical*, *split:coordination* eta *insert:missing main verb* aipatzen dituzte.

7.4 Etiketatzearen emaitzak

Atal honetan corpusean egindako lerrokatzeen eta eragiketen intzidentzien emaitzak emango ditugu. Hurbilpenen arteko konfrontazioa ere hemen jasoko dugu. Aurkeztuko ditugun emaitzak beste hizkuntzetakoekin erkatzen saiatu gara.

Testua	**Bertsioa**	**Esaldi-kopurua**	**Hitz-kopurua**
Bernoulli (Teknologia)	jatorrizkoa	89	1.446
	itzultzailea	123	1.472
	irakaslea	105	1.253
Etxeko (Medikuntza)	jatorrizkoa	70	1.535
	itzultzailea	84	1.611
	irakaslea	75	1.608
Exoplanetak (Giza Zientziak)	jatorrizkoa	68	1.512
	itzultzailea	75	1.608
	irakaslea	96	1.258
Guztira	jatorrizkoa	227	4.493
	itzultzailea	282	4.691
	irakaslea	276	4.119
Guztira	corpusa	785	13.303

7.15 taula – Esaldi- eta hitz-kopurua jatorrizko testuetan eta testu sinplifikatuetan

Etiketatzearen emaitzak azaltzen hasi aurretik, begiratu bat egin diezaiogun sortu dugun corpus laginaren ezaugarriei. Hiru testu erabili ditugu fase honetan: “Bernoulli” (Teknologia), “Etxeko” (Medikuntza), eta “Exoplanetak” (Giza Zientziak eta Historia). 7.15 taulan testu horien hitzen eta

esaldien kopuruak ikus daitezke. Corpus guztia kontuan izanda, 785 esaldi dauzkagu eta 13.303 hitz.

Esaldi-kopuruei erreparatuz, sinplifikatutako testu guztietan esaldi gehiago aurkitzen ditugu jatorrizkoetan baino. Hitzen kopurua kontuan hartuz, itzultzaileak sinplifikatutako testuetan hitz gehiago daude. Joera hori irakasleak sinplifikatutako batean ere ("Etxeko") ikusi dugu. Irakasleak sinplifikatutako beste bi testuetan, aldiz, hitz-kopurua jaitsi egin da.

Erdal hizkuntzetako corpusei begiratu bat eginez, gurean bezala, eskuz sinplifikatu direnak aztertuko ditugu hemen; hau da, jatorrizko testu batetik sortutako testu sinplifikatuak biltzen dituzten corpusak. Ingelesez, 104 artikulu bildu dituzte bakoitzari dagokion bertsio murriztuarekin[3] (*abridged*). Jatorrizkoan 2.539 esaldi jaso dituzte eta sinplifikatutakoan 2.459 daude; hitzak, berriz, 41.982 dira jatorrizkoan eta 29.584 sinplifikatutakoan (Petersen eta Ostendorf, 2007).

Brasilgo portugeserako sortutako corpusean bi sinplifikazio-maila bereizi dituzte; hala, corpusean, jatorrizko testuak, maila naturalean sinplifikatutakoak eta maila absolutuan sinplifikatutakoak aurkitu ditugu. Jatorrizko testuen esaldi-kopurua 2.116 da, naturalena 3.104 eta absolutuena 3.537; hitzen kopuruak, berriz, jatorrizkoan 41.897 dira, 43.013 naturalean eta 43.676 absolutuan (Caseli *et al.*, 2009).

Gaztelaniaz, esaldien lerrokatze automatikoa egiteko erabiltzen duten corpus laginean (egileen arabera, erabilgarri dutena) jatorrizko 110 esaldi daude eta 145 sinplifikatu. Hitzak, aldiz, 2.456 dira jatorrizkoan eta sinplifikatutakoan 1.840. Berez, *Simplext* corpusa 200 dokumentuko testu-bilduma dela aipatzen dute Bott eta Saggion-ek (2011). Beste lan batean, berriz, 37 artikulu pare erabili dituzte Štajner *et al.*ek (2013). FIRST izeneko 25 dokumentuko corpusak, berriz, 330 esaldi ditu (Štajner, 2015).

Danieraz, 3.701 testu pareekin osatu dute corpusa. Lerrokatuta, 48.186 esaldi dituzte jatorrizko testuen atalean eta 62.365 esaldi sinplifikatutakoen atalean (Klerke eta Søgaard, 2012). Lerrokatu gabe esaldi gehiago dituzte.

Italieraz, corpusa bi azpicorpusez osatuta dago: ikuspegi estrukturala biltzen duen *Terence* eta intuitiboa biltzen duen *Teacher*. Ikuspegi estruturalarekin sinplifikatutakoan, jatorrizkoan 1.036 esaldi daude eta sinplifikatutako corpusean 1.060. Ikuspegi intuitiboan, berriz, 24 artikulu pare (Brunato *et al.*, 2015).

[3]Corpus hau testuen sinplifikazioa aztertzeko erabili dutenez aipatzen dugu hemen, gure sinplifikazioaren ikuspegia ez den arren.

Datu horiek guztiak guk osatu dugun corpusarekin alderatuta esan behar da nabarmenki txikiagoa dela gurea. Salbuespen bakarra gaztelaniazko lagina da. Hitz- eta esaldi-kopuruak direla-eta, esan behar dugu, euskarazko corpusean bezala, portugesezkoan, danierazkoan, gaztelaniazkoan eta italierazko *Terence* corpusean esaldi-kopurua igo egiten dela sinplifikatutako bertsioetan. Hitz-kopuruak, berriz, portugesez gora egiten du, baina gaztelaniaz behera. Ingelesezko corpusean, hitz- eta esaldi-kopuruek behera egiten dute. Kopuru horiek guztiak 7.16 taulan ikus daitezke.

Corpusa/Hizkuntza	**Art./-Dok.**	**Esaldiak**		**Hitzak**	
		Jat.	**Sinp.**	**Jat.**	**Sinp.**
Ingelesa Petersen eta Ostendorf (2007)	104	2.539	2.459	41.982	29.584
Brasilgo portugesa Caseli *et al.* (2009)	-	2.116	3.104 (nat.) 3.537 (bor.)	41.897	43.013 (nat.) 43.676 (bor.)
Gaztelania Bott eta Saggion (2011)	-	110	145	2.456	1.840
Daniera Klerke eta Søgaard (2012)	3.701	48.186	62.365	-	-
Italiera *Terence* Brunato *et al.* (2015)	-	1.036	1.060	-	-
Italiera *Teacher* Brunato *et al.* (2015)	24	-	-	-	-

7.16 taula – Erdal hizkuntzetako corpusetan dauden kopuruak

7.4.1 Lerrokatzeak

Testuan eragiketak etiketatu ahal izateko, testuak lerrokatu egin behar dira. Hau da, jatorrizkoaren x-esaldiari sinplifikatutako testuen zein esaldi dagokion edo dagozkion adierazi behar da, eragiketen etiketatzea eraginkorra izan dadin. Lerrokatze horiek zein eskalatan gauzatu diren 7.17 taulan ikus dezakegu. Taulan azaltzen dugun eskalaren bitartez, jatorrizko testu-

ko esaldi bakoitzeko sinplifikatutako testuko zenbat esaldi lerrokatzen diren adierazi dugu. Adibidez, 1:1 eskalan jatorrizko esaldi bati esaldi sinplifikatu bat dagokio eta 1:2 eskalan jatorrizko bati bi sinplifikatu dagozkio.

Eskala	**Itzultzailea**	**Irakaslea**
1:1	76,21	73,25
1:2	18,50	19,74
1:3	3,52	4,39
2:1	0,88	0,44
Bestelakoak	0,88	2,19

7.17 taula – Lerrokatzeen emaitzak

Bi sinplifikatzaileen esaldi gehienak 1:1 eskalan lerrokatu ditugu, hau da, jatorrizko esaldi bakoitzeko esaldi sinplifikatu bat eman dute. Itzultzaileak % 76,21ean egin du lerrokatze hori eta irakasleak % 73,25ean. Jatorrizko esaldia bitan banatzea (1:2 lerrokatzea) izan da gehien egin den bigarren parekaketa; itzultzaileak % 18,50 esaldietan egin du eta irakasleak % 19,74etan. Jatorrizkoa hiru esalditan banatzea (1:3 lerrokatzea) gutxiagotan egin da; itzultzaileak esaldien % 3,52an eta irakasleak % 4,39an. Jatorrizko bi esaldi bat bihurtzea (2:1 lerrokatzea), itzultzaileak % 0,88an egin du eta irakasleak % 0,44an. Bestelako lerrokatzeak dira hiru esaldi baino gehiagotan banatu diren esaldiak eta esaldi erdiak lerrokatu diren esaldiak. Itzultzaileak % 0,88an egin ditu eta irakasleak % 2,19an. Kopuruei, oro har, begiratzen badiegu nabarmentzekoa da bi sinplifikatzaileek antzeko maiztasunarekin egin dituztela lerrokatzeak.

Lerrokatzeak beste hizkuntzetan nolakoak izan diren ere aztertu dugu. Ingelesez (Petersen eta Ostendorf, 2007), italieraz (Brunato *et al.*, 2015) eta gaztelaniaz ere (Štajner, 2015) 1:1 izan da lerrokatze ohikoena. Bigarren erabiliena, aldiz, ingelesez 1:0 izan da eta italieraz 2:1 ikuspegi intuitiboan eta 1:2 ikuspegi estrukturalean. Gaztelaniaz ere, bigarren ohikoena 1:n (jatorrizko esaldia esaldi bat baino gehiagotan banatzea, ez dute zehazten zenbatetan) izan da eta ondoren 1:0 (bi corpusetan).

7.4.2 Eragiketen intzidentzia

Azpiatal honetan etiketatzearen emaitzak jasoko ditugu eragiketen intzidentzia ikusteko; hau da, sinplifikatzaile bakoitzak makroeragiketa eta eragiketa

bakoitza zein maiztasunarekin erabili duen ikusiko dugu.

Makroeragiketei dagokienez, 7.18 taulan ikus daitekeen bezala, itzultzaileak sinplifikatutako testuetan, makroeragiketarik erabiliena transformazioa (% 24,92) da, baina banaketa (% 23,55) ere oso gertu dago. Bateratzea (% 0,40) eta bestelakoak (% 0,10) izan dira gutxien erabili dituen makroeragiketak. Irakasleak sinplifikatutako testuetan ere makroeragiketarik erabiliena transformazioa (% 33,62) da, baina bigarren erabiliena ezabatzea (% 20,78) da. Bigarren makroeragiketa horren erabilerarekin ulertzen dugu zergatik irakaslearen bi testuek jatorrizkoek baino hitz gutxiago dauzkaten. Gutxien erabili duena bateratzea (% 0,10) izan da eta ez du bestelakorik erabili.

Makroeragiketa	**Itzultzailea**	**Irakaslea**
Transformazioa	24,92	33,62
Banaketa	23,55	12,30
Txertaketa	21,88	18,61
Ezabatzea	17,66	20,78
Hurrenkera-aldaketa	7,95	8,27
Eragiketarik ez	3,53	6,20
Bateratzea	0,40	0,22
Bestelakoa	0,10	0,00

7.18 taula – Makroeragiketen maiztasunak

Horretaz gain, irakasleak transformazioak maizago erabili ditu (% 33,62) itzultzaileak (% 24,92) baino. Sinplifikatu gabe utzitako esaldiak ere gehiago izan dira; izan ere, irakasleak % 6,20an ez du eragiketarik egin eta itzultzaileak, aldiz, % 3,53an. Hurrenkera-aldaketa, txertaketa eta ezabatzea makroeragiketen ehunekoak nahiko antzekoak dira. Banaketak, berriz, itzultzaileak (% 23,55) gehiagotan egin ditu irakasleak baino (% 12,30).

Bietan transformazioa izatea makroeragiketarik erabiliena ez da harritzekoa; izan ere, eragiketa ugari bere gain hartzeaz gain testuak sinplifikatzean berridazketak egiten dira, eta, jakina denez, berridazketa horietan batez ere transformazioak egiten dira.

Transformazioen analisiarekin hasiko dugu makroeragiketen banan-banako azterketa. Makroeragiketa horretan *Sub2Main* (mendeko perpaus bat perpaus nagusi bihurtzea) izan da eragiketarik erabiliena itzultzailearen testuetan. Eragiketa hori transformazioen % 48,50 da. Irakaslearen testuetan,

aldiz, eragiketarik erabiliena *Reform* izan da, transformazioen % 19,09rekin. Datu horiekin ikusten da itzultzaileak testuak transformatzean, mendeko perpausak nagusi bihurtzeko joera handia izan duela eta irakasleak, aldiz, eragiketa-barietate gehiago erabili duela.

Transformazio-eragiketak motaren arabera sailkatzen baditugu (7.19 taula), ikus dezakegu bi sinplifikatzaileek gehien erabili dituzten transformazioak sintaktikoak direla. Gutxien erabilitakoak, aldiz, zuzenketak izan dira.

Transformazio-mota	Itzultzailea	Irakaslea
Sintaktikoak	41,34	33,01
Morfologikoak	22,05	19,09
Bestelakoak	17,57	19,74
Diskurtsokoak	14,96	15,86
Lexikalak	6,70	11,03
Zuzenketak	0,39	1,29

7.19 taula – Transformazio-moten maiztasunak

Bietan transformazio sintaktikoa gehien erabilitakoa izateak azpimarratzen du gure ustez sintaxiak testuak sinplifikatzean duen garrantzia. Transformazio sintaktikoak eta lexikoak alde batera utzita, gainontzeko motetan ez dago puntu askoko ezberdintasunik. Itzultzaileak sintaxiari garrantzi handiagoa eman dio irakasleak baino (zortzi puntu baino gehiagoko aldea) eta irakasleak lexikoari (lau puntu baino gehiagoko aldea). Halaber, nabarmentzekoa da transformazio morfologikoen indarra. Bestelako transformazioek ere kopuru handia dute eta horiek etorkizuneko analisi baten beharra adierazten dute, azpisailkapenik egin daitekeen ikusteko.

Banaketak analizatuz, indarraren arabera, itzultzaileak % 74,06ko maiztasunarekin erabili duen banaketa-mota *soft* edo leuna da. Hau da, egin diren perpausen banaketetatik ia hiru laurden puntu eta komarekin egin dira. Irakasleak, aldiz, banaketa gehienak gogor egin ditu, % 69,03ko maiztasunarekin; leunek, berriz, % 30,97ko maiztasuna dute. Ikus daitekeenez, bi hurbilpenak zeharo ezberdinak dira puntu honetan.

Banatutako fenomenoei erreparatuz (7.20 taula), perpaus koordinatuak izan dira gehien banatutako fenomenoa (itzultzaileak % 39,17 eta irakasleak % 45,13) eta ondoren perpaus adberbialak (itzultzaileak % 19,16 eta irakasleak % 16,81).

Banaketa (fenomenoa)	Itzultzailea	Irakaslea
Koordinazioa	39,17	45,13
Perpaus adberbialak	19,16	16,81
Perpaus erlatiboak	16,25	11,50
Aposizioak/egitura parentetikoak	10,83	7,96
Osagarriak	7,50	0,00
Postposizioak	3,75	3,54
Bestelakoak	3,33	15,05

7.20 taula – Fenomenoen araberako banaketen maiztasunak

Itzultzaileak sinplifikatutako testuetan, % 10etik gora ditugu perpaus erlatiboak (% 16,25) eta aposizio-sintagmak (% 10,83); % 10etik behera, berriz, perpaus osagarriak (% 7,50), postposizioak (% 3,75) eta bestelakoak (% 3,33). Irakasleak sinplifikatutako testuetan, % 10etik gora ditugu bestelakoak (zerrendatze edo enumerazioak eta estilo mistoa) (% 15,05) eta erlatibozko perpausak (% 11,50); % 10etik behera, berriz, aposizio-sintagmak (% 7,96) eta postposizioak (% 3,54). Nabarmentzekoa da irakasleak ez dituela perpaus osagarriak banatu.

Banatu diren perpaus adberbialen artean (7.21 taula), itzultzailearen testuetan, % 20tik gora baldintza-perpausak (% 23,91) eta kausa-perpausak (% 21,74) dira; % 10etik gora, modu-perpausak (% 17,39), denbora-perpausak (% 13,04) eta kontzesio-perpausak (% 10,87) eta % 10etik behera, helburu-perpausak (% 6,52) eta konparazio-perpausak (% 6,52). Irakaslearen testuetan, aldiz, % 20tik gora kausa-perpausak (% 42,11) eta denbora-perpausak (% 26,32) dira; % 10etik gora, kontzesio-perpausak (% 15,79) eta helburu-perpausak (% 10,52) eta % 10etik behera, modu-perpausak (% 5,23). Konparazio-perpausik ez du banatu. Emaitza horiek ikusita, zaila egiten da perpaus adberbialen intzidentzietan patroi komunik aurkitzea.

Behin banaketak zein fenomenotan izan diren aztertu eta gero, 7.22 taulan, jatorrizko testuak kontuan izanda, ehuneko zenbatetan egin diren ikusiko dugu. Hau da, jatorrizko testuetan fenomeno bakoitzaren agerpenak kontatu ditugu[4] (bigarren zutabea) eta horietatik sinplifikatzaile bakoitzak egin dituen banaketen ehunekoa adierazi dugu. Datu hauen helburua da egindako banaketen ehunekoen eta banatutako fenomeno ehunekoen arteko kontrastea

[4]Kontaketak egiteko ErreXail erabili dugu. Sistemaren ezaugarriak direla-eta ezin izan ditugu konparazio-perpausak automatikoki monitorizatu.

Banaketa (adberbialak)	Itzultzailea	Irakaslea
Baldintza	23,91	0,00
Kausa	21,74	42,11
Modua	17,39	5,23
Denbora	13,04	26,32
Kontzesioa	10,87	15,79
Helburua	6,52	10,52
Konparazioa	6,52	0,00

7.21 taula – Banatutako perpaus adberbialen maiztasunak

ikustea.

Mendeko mota	Kopurua (jat.)	Banatutakoak (itzul.)	Banatutakoak (irak.)
Osagarria	162	11,11	0,00
Modua	69	11,59	1,45
Erlatiboa	57	66,67	22,81
Baldintza	57	19,30	0,00
Denbora	34	17,65	14,71
Kausa	23	43,48	34,78
Helburua	20	15,00	10,00
Modu/denbora	17	0,00	0,00
Kontzesioa	5	100,00	60,00

7.22 taula – Banatutako mendeko perpausen proportzioa

Itzultzaileak gehien banatu dituen fenomenoak kontzesio-perpausak (% 100,00), erlatibozko perpausak (% 66,67) eta kausa-perpausak (% 43,48) izan dira, eta gutxien, aldiz, modu-perpausak (% 11,60) eta perpaus osagarriak (% 11,11). Baldintza- (% 19,30) eta denbora- (% 17,65) perpausak antzera banatu dira. Modu/denbora eta helburu-perpausak ez ditu banatu.

Irakasleak gehien banatu dituenak ere kontzesio-perpausak (% 60,00) izan dira, ondoren, kausa-perpausak (% 34,78) eta erlatibozko perpausak (% 22,81). Gutxien banatutakoak denbora-perpausak, (% 14,71), helburu-perpausak (% 10,00) eta modu-perpausak (% 1,45) izan dira. Banatu ez dituenak perpaus osagarriak, baldintza-perpausak eta modu/denbora-perpausak

izan dira.

Bi hurbilpenak parekatzen baditugu, hiru mota banatuenak (kontzesio-, kausa- eta erlatibozko perpausak) berdinak dira bietan, hiruren arteko hurrenkera aldatzen den arren. Denbora-perpausen proportzioa ere antzekoa da eta biek ez dituzte modu/denbora-perpausak banatu. Gainontzeko perpaus motekin, aldiz, estrategia ezberdinak baliatu dituzte.

Insert edo elementuen txertaketa da maiz erabili den beste makroeragiketa bat. Etiketatzean bereizi ditugun hiru txertaketa moten emaitzak 7.23 taulan ikus daitezke.

Txertaketa-motak	**Itzultzailea**	**Irakaslea**
Beharrezkoak ez direnak	44,39	57,89
Esaldi berrietan	42,15	30,99
Elidítutakoak	13,45	11,11

7.23 taula – Txerkaketen emaitzak (multzoka)

Beharrezkoak ez diren txertaketak % 44,39 izan dira itzultzailearen testuetan eta % 57,89 irakaslearen testuetan. Esaldi berriak sortzeko txertatu behar izan diren elementuak % 42,15 izan dira itzultzailearentzat eta % 30,99 irakaslearentzat. Elidítutako elementuen berreskuratzea gutxiagotan egin dute bi sinplifikatzaileek (itzultzaileak % 13,45 eta irakasleak % 11,11), ziurrenik fenomeno hori ere ez delako hainbestetan gertatzen, baina automatikoki detektatzeko eta konfirmatzeko zaila egin zaigu.

Bi hurbilpenetan egin diren eragiketen rankinga bera bada ere, irakasleak 13 puntu baino gehiagotan txertatu ditu beharrezkoak ez diren elementuak. Itzultzaileak, aldiz, 11 puntu baino gehiagotan txertatu ditu elementuak esaldi berrietan. Elidítutako elementuen txertaketan ez dago ezberdintasun handirik.

Delete makroeragiketan, bi ezabatze-mota bereizi ditugu informazioari dagokion tratamenduaren arabera. Informazioa ezabatzen dutenak itzultzailearen testuetan % 25,56 dira, eta irakaslearenetan % 30,37. Hitz funtzionalak, berriz, % 74,44tan ezabatu ditu itzultzaileak eta % 69,36tan irakasleak. Hau da, bi hurbilpenetan hitz funtzionalen ezabatzeak izan dira nagusi. Datu horiek 7.24 taulan ikus daitezke.

Informazioa ezabatua izan den kasuetan azterketa sakonagoa egin behar da, multzo irregularra eta zabala baita. Analisi sakonago horretan ikusiko dira nolakoak izan diren eta ea azpisailkapenik egin daitekeen. Hitz/token

Ezabatze-motak	Itzultzailea	Irakaslea
Informazioa	25,56	30,37
Hitz funtzionalak	74,44	69,36

7.24 taula – Ezabatzeen emaitza

funtzionalak, aldiz, multzo itxia dira eta 7.25 taulan jaso ditugu gehien ezabatu direnen emaitzak (etiketatzean mota bereizi behar zen).

Hitz funtzionalen ezabatzeak	Itzultzailea	Irakaslea
Juntagailuak	54,48	33,08
Puntuazio-markak	23,88	34,59
Diskurtso-markatzaileak	14,93	24,06
Bestelakoak	6,71	8,27

7.25 taula – Hitz funtzionalen ezabatzeen emaitzak

Ezabatu diren hitz/token funtzional gehienak juntagailuak, puntuazio-markak eta diskurtso-markatzaileak izan dira bi hurbilpenetan, proportzioak eta erabilienen hurrenkera ezberdina den arren. Itzultzaileak juntagailuak % 54,48an, puntuazio-markak % 23,88an eta diskurtso-markatzaileak % 14,93-an ezabatu ditu eta irakasleak puntuazio-markak % 34,59an, juntagailuak % 33,08an eta diskurtso-markatzaileak % 24,06an ezabatu ditu.

Hurrenkera-aldaketen eragiketei helduz, 7.26 taulan ikus dezakegu sintagmen posizio-aldaketa izan dela eragiketarik erabiliena bi hurbilpenetan (itzultzaileak % 43,20 eta irakasleak % 78,95). Sintagmak tokiz aldatzea eta esaldi berrietan sartzea izan da itzultzailearen bigarren eragiketarik erabiliena (% 41,98) eta gutxiagotan aldatu du perpausen hurrenkera (% 13,58). Oso kasu gutxitan aldatu du aditz laguntzaileen hurrenkera (% 1,23). Irakasleak, berriz, perpausak tokiz aldatu ditu bigarren postuan (% 13,16) eta sintagmak esaldi berrietara pasatzea izan da gutxien egin duena (% 7,89); ez ditu aditz laguntzaileak mugitu. Nabarmentzekoa da perpausen hurrenkera-aldaketen ehunekoa oso antzekoa dela bi hurbilpenetan.

Hurrenkera-aldaketa horiek analizatzen jarraitzea interesgarria izango da etorkizunean. Sintagmen hurrenkera-aldaketetan, sintagma horiek hurrenkera kanonikoa edo besteren bat betetzeko mugitu diren begiratu beharko genuke eta perpausen hurrenkera-aldaketak guk corpusean aurkitu ditugun joerekin bat datozen konprobatu beharko genuke.

Hurrenkera-aldaketa motak	Itzultzailea	Irakaslea
Sintagmak	43,20	78,95
Sintagmak esaldi berrietan	41,98	7,89
Perpausak	13,58	13,16
Aditz laguntzaileak	1,23	0,00

7.26 taula – Hurrenkera aldaketen emaitzak

Etiketatze-eskeman definitu ditugun gainontzeko makroeragiketak (eragiketarik eza, bateratzea eta bestelakoak) ez dira % 5era iristen itzultzailearen testuetan, ez eta hiruren arteko batuketa eginda ere. Irakaslearen testuetan, aldiz, eragiketarik eza % 6,20an gertatu da eta bateratzea % 0,22an. Makroeragiketa horiei dagozkien emaitzak 7.27 taulan ikus daitezke.

Gainontzeko makroeragiketak	Itzultzailea	Irakaslea
Eragiketarik eza	3,53	6,20
Bateratzea	0,40	0,22
Bestelakoak	0,10	0,00

7.27 taula – Gainontzeko makroeragiketak emaitzak

Eragiketarik egin ez den esaldietan analisi sakonago bat egin behar da, beraien ezaugarriak ezagutzeko. Izan ere, noiz ez den sinplifikatu behar jakitea beharrezkoa da testuen sinplifikazioan. Bateratze gutxi aurkitzea ez zaigu arraroa egiten eragiketa horrek gure ustez laburpenarekin zerikusi handiagoa duelako. Bestelako eragiketarekin esaldi bakarra etiketatu da. Esaldi horretan makroeragiketa batek baino gehiagok hartzen dute parte eta zaila izan da hori sailkatzea.

Hurbilpenen konfrontazioaren laburpena

Bi hurbilpenen konfrontazioa laburbiltzeko asmoz, bi hurbilpenek amankomunean izan dituzten emaitzak adieraziko ditugu. Makroeragiketarik erabiliena bi kasuetan transformazioa izan da, eta transformazio-mota erabiliena sintaktikoa. Perpausen banaketari dagokionez, biek banatu dituzte koordinazioak eta perpaus adberbialak gehien. Mendeko perpausetan, jatorrizko testuetan dauden esaldiak kontuan izanda biek gehien banatu dituzten per-

pausak kontzesio-, kausa-perpausak eta perpaus erlatiboak izan dira. Zuzenketaren beharra ere bi hurbilpenetan ikusi da, gutxi bada ere.

Ondorio gisa, emaitza horiek egiaztatzen dute orain arte aipatu ditugun puntu horiek guk proposatzen dugunarekin bat datozela. Alegia, guk sinplifikazio sintaktikoa aukeratu dugu, eta batez ere koordinazioari eta perpaus adberbialei eman diegu garrantzia banaketak eta berreraikitzeak egiterakoan. Zuzenketaren beharra ere azpimarratu dugu gure proposamenean.

Komun izan dituzten beste eragiketen artean, beharrezkoak ez diren txertaketak aurkitu ditugu. Fenomeno hori gurean etorkizuneko lan bezala aipatu izan dugu; lan hori gauzatuko duen korreferentzia sistema prest dagoenean edo beste iturrietatik (adibidez, *Wikipedia*) edukiak txertatzeko gai garenean. Ezabatze funtzionalak, berriz, guk kategoria bezala tratatu ez ditugun arren, egia da ezabatze horiek gure erregeletan aurki daitezkeela (koordinazioa edo puntuazio-markak, adibidez), baina beste batzuk (diskurtso-markatzaileak ezabatzea, esaterako) ez. Sintagmen hurrenkera-aldaketak direla-eta, hurrenkera kanonikoa jarraitzea proposatu dugun arren, mementoz jatorrizko esaldian dauden hurrenkerei eutsiko diegu. Izan ere, perpaus adberbialekin egin dugu bezala, corpus-azterketa bat beharko luke, eta, horretaz gain, hemen egin diren mugimenduak sakonago analizatu beharko genituzke.

Beraz, beste ondorio bat da bi hurbilpenetan komun aurkitu ditugun ezaugarriak etorkizunerako pentsatuta daukagunarekin ere bat datozela. Hala ere, aipatu nahi dugu guk orain arte sistematikoki aztertu ez ditugun transformazio morfologikoek izan dutela intzidentzia, eta uste dugu etorkizunean aztertu beharko ditugula. Dena dela, hurbilpen bakoitzetik gehiago ikas daiteke eta horrek lagunduko digu helburu-taldeen araberako sinplifikazioak zehazten.

Testuak eskuz sinplifikatzeko eragiketen zerrendan hurrengo puntuak hartu ditugu kontuan: sintaxi-mailako transformazioa egin, perpaus koordinatuen, kontzesiboen, kausalen eta erlatiboen banaketan erreparatuz; beharrezkoa ez den informazioa ere gehitu, elidi tutako subjektuak, objektuak eta abar berreskuratuz.

Atal honekin amaitzeko, aipatu nahi dugu 7.3 azpiatalean aurkeztutako etiketatze-eskema baliagarria izan dela testuak etiketatzeko, analizatzeko eta sinplifikatzaileen arteko konparazioak egiteko. Ziur gaude etiketatze-eskema osatu eta sakondu daitekeela, eta lan hori eta hortik eratorritako analisia beste fase baterako utziko dugu.

Beste hizkuntzekin erkatuz

Etiketatze-eskemen konparazioa egiterakoan ikusi dugu makroeragiketa mailan (klase edo lehenengo dimentsioan) italierazko eta gaztelaniako eskemak gurearen oso antzekoak direla. Hortaz, azpiatal honetan etiketatzearen maila horretako emaitzak beste bi hizkuntza horiekin erkatuko ditugu eta hurrengo mailakoak ere saiatuko gara erkatzen. Portugeseko corpusean aurkitutako eragiketen emaitzekin ere egingo dugu konparazioa. Konparazio hori zailagoa da, batetik, etiketatze-eskemarik ez dagoelako eta, bestetik, emaitzak bereizten dituzten sinplifikazio-mailen arabera ematen dituztelako. Konparazioa egiteko datuak 7.28 taulan jaso ditugu; datu horiek Bott eta Saggion-en (2014) eta Brunato *et al.*en (2015) lanetatik atera ditugu.

Gaztelaniaz gehien erabili den makroeragiketa transformazioa izan da, bigarrena ezabatzea eta hirugarrena txertaketa (Bott eta Saggion, 2014). Italierako bi hurbilpenetan ere hiru eragiketa horiek aurkitu ditugu erabilienen artean (Brunato *et al.*, 2015). Eta hiru makroeragiketa horiek ere bat datoz gure irakaslearen ikuspegiarekin. Itzultzaileak, ordea, banaketak bigarren postuan egin ditu (transformazioak lehenengoan eta txertaketak hirugarrenean, besteetan bezala).

Kopuruak begiratuta, italieraz eta euskaraz egin den hurrenkera-aldaketa eragiketen proportzioa antzekoa da eta bi azpicorpusen (*Terence* eta *Teacher*) kasuan txertaketak ere antzekoak dira. Gaztelaniaz, berriz, proportzioa baxuagoa da bi eragiketetan. Gutxi erabili diren makroeragiketen artean bateratzea dugu, zeina hiru hizkuntzetan eta hurbilpen ezberdinetan oso gutxi agertzen den. Nabarmentzekoa da ere euskaraz perpausen banaketa askoz ere maizago erabili dela.

Transformazio-motak aztertuz, gaztelaniaz eta italieraz gehien egin direnak transformazio lexikalak izan dira. Portugeseko corpusean ere ordezkapen lexikala izan da gehien egin den eragiketa jatorrizko testutik sinplifikazio naturala egiterakoan (Caseli *et al.*, 2009). Gurean, aldiz, sintaktikoak izan dira gehien erabilitakoak. Perpausen banaketari helduz, gaztelaniaz, gurean bezala, koordinazioa izan da gehien banatutako fenomenoa. Gainontzeko datuak erkatzea zailagoa egiten zaigu eta datu gehiagoren beharra ikusten dugu. Hala ere, ondoriozta daiteke transformazioak oso garrantzitsuak direla sinplifikazioan.

Atal honi bukaera emanez, ikusi dugu jatorrizko testu batetik testu sinplifika-

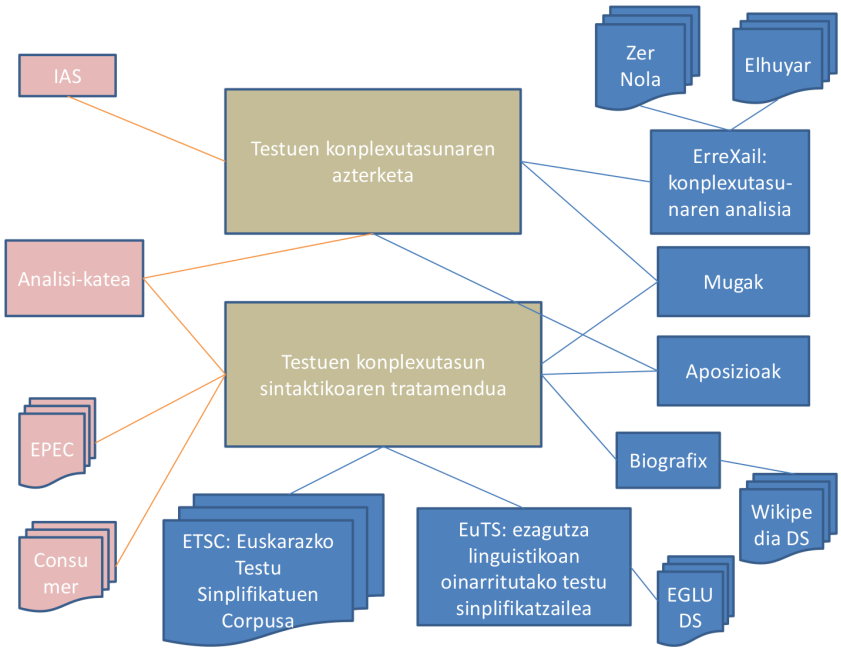
Makroeragiketa	Italiera		Gaztelania	Euskara	
	Terence	Teacher		Itzultzailea	Irakaslea
Transformation	48,18	47,76	39,02	24,92	33,62
Split	1,71	2,06	12,20	23,55	12,30
Insert	18,72	15,66	12,60	21,88	18,61
Delete	21,94	25,32	24,80	17,66	20,78
Reordering	8,65	7,89	2,85	7,95	8,27
No_operation	-	-	-	3,53	6,20
Merge	0,81	1,30	0,81	0,40	0,22
Other	-	-	-	0,10	0,00

7.28 taula – Makroeragiketen konparazioa hizkuntza artean

tu bat lortzeko modu bat baino gehiago dagoen arren, hizkuntza eta hurbilpen ezberdinetan antzeko eragiketak gauzatzen direla. Datu horietan oinarrituz pentsatzen dugu hizkuntza arteko sinplifikazio-erregela orokorrak egin daitezkeela, adibidez transformazioei lehentasuna emanez.

7.5 Laburpena

Kapitulu honetan euskarazko testu sinplifikatuen corpusa nola osatu eta etiketatu dugun azaldu dugu. Etiketatze hori egiteko, etiketatze-eskema sortu dugu corpusean egindako analisietan oinarrituz eta erdal hizkuntzetako etiketatze-eskemak aztertuz. Etiketatze hori baliatuta jatorrizko testuetatik testu sinplifikatuak sortzeko makroeragiketen eta eragiketen emaitzak eman ditugu eta bi hurbilpenen arteko konparazioa egin dugu. Bi hurbilpen horiek gure proposamenarekin eta beste hizkuntzetako lanekin ere erkatu ditugu. Honenbestez, 7.2 irudian baliabideen eta ekarpenen egoeran ETSC corpusa gehitu dugu.



7.2 irudia – Tesian erabilitako baliabideak eta tresnak, eta egindako ekarpenak

8 Ondorioak eta etorkizuneko lanak

8.1 Sarrera

Tesi-lan honen motibazioa HPko tresna aurreratuetan esaldi luzeek eta konplexuek sortzen dituzten arazoak ebaztea eta euskara ikasten ari direnei testu sinpleagoak edo errazagoak eskaintzea izan da, Ixa taldearen tresnen berrerabilpena/egokitzapena aztertuz. Horretarako, bi ikerketa-lerro aztertu ditugu: testuaren konplexutasunaren analisia eta testuen sinplifikazio automatikoa. Ikerketa-lerro horiek euskarazko prozesamendura ekarri ditugu beste hizkuntzetan egin diren lanetan oinarrituta. Zehazki, idatzizko testu konplexuen azterketa linguistikoa egin dugu eta sinplifikazioa automatikoki gauzatzeko sistemaren oinarri linguistiko-konputazionalak finkatu ditugu.

Gure proiektuan bi helburu planteatu ditugu: i) alde linguistikotik, testuen konplexutasuna aztertzea eta sinplifikazio-proposamenak egitea eta, ii) tratamendu konputazionaletik, testuen konplexutasuna neurtzea eta testuak automatikoki sinplifikatuko dituen sistema informazio linguistikoaz hornitzea eta inplementatzea. Hurrengo ataletan, helburu horiek lortzeko egin ditugun ekarpenak eta aurreikusten ditugun etorkizuneko lanak azalduko ditugu.

8.2 Ekarpenak

Tesi-lan honetan egindako ekarpen nagusiak sarrera-kapituluan aipatutako ikerketa-galdera multzoen eta ikerketa-lerroen arabera adieraziko ditugu.

8.2.1 Konplexutasunaren azterketa eta testuen konplexutasunaren analisia

Testu bat konplexutzat hartzeko irizpideak finkatu eta **egitura konplexuak definitu** ditugu. Euskarazko egitura konplexuak zeintzuk diren definitzeko, beste hizkuntzetan egindako lanetan oinarritu gara eskuzko azterketa linguistikoari abiapuntua emateko. Azterketa linguistiko horretan egitura konplexutzat hartu ditugu perpaus koordinatuak, perpaus osagarriak, perpaus erlatiboak, perpaus adberbialak, aposizio-sintagmak eta egitura parentetikoak. Horietaz gain, usteak edo adierazpenak adierazten dituzten postposizio-sintagmak analizatu ditugu eta egitura horiek guztiak sinplifikagarriak izateko, aditzaz gain gutxienez beste bi argumentuko edo adjuntuko luzera minimoa izan behar dutela zehaztu dugu.

Testuen konplexutasunaren analisi automatikoan **ErreXail sistema sortu** dugu. ErreXail sistemak 94 ezaugarri linguistikotan oinarrituta, SMO sailkatzailea (Platt, 1998) erabiltzen du testuak sinpleak ala konplexuak diren adierazteko. Horretaz gain, ErreXail sistemaren bidez testu konplexuak eta sinpleak bereizten laguntzen dituzten ezaugarri esanguratsuenenen zerrenda lortu dugu estatistikan oinarrituta.

Ikerketa-galdera multzo hori 3. eta 5. kapituluetan erantzun dugu.

8.2.2 Konplexutasunaren tratamendua eta testuen sinplifikazio automatikoa

Testuen sinplifikazio automatikoa gauzatzeko, **EuTS sistemaren diseinu linguistikoa egin** dugu. EuTS sistema konplexutasunaren azterketa linguistikoan oinarritutako erregelak aplikatuko dituen sistema da. EuTS sistemak bi sinplifikazio-mota egiten ditu: ordezkapen sintaktikoen sinplifikazioa eta sinplifikazio sintaktikoa. Ordezkapen sintaktikoen sinplifikazioan maiztasun gutxiko egitura sintaktikoak maiztasun handiagoa dutenekin ordezkatzen dira eta sinplifikazio sintaktikoan, berriz, egitura konplexu horiek kentzen dira egitura-aldaketak eginez. Sinplifikazio-mota horiek egiteko bost eragiketa zehaztu ditugu: ordezkapen sintaktikoen sinplifikazioan, i) azaleko ordezkapen sintaktikoak eta sinplifikazio sintaktikoan, i) banaketa, ii) esaldien berreraikitzea, iii) esaldien ordenatzea eta iv) esaldien zuzenketa eta egokitzapena. Esaldiak berreraikitzean txertatze-elementuak monotonoak izan ez daitezen, txertatze-elementu alternatiboak proposatu ditugu.

EuTS sistemaren helburu-taldea nahiko irekia da, baina euskara ikasten

ari direnentzat eta HPko aplikazio aurreratuentzat hiru sinplifikazio-maila definitu ditugu: azaleko sinplifikazio sintaktikoa, sinplifikazio naturala eta sinplifikazio absolutua. Lehena, ordezkapen sintaktikoen sinplifikazioa, euskara-maila aurreratua duten ikasleei, baina egitura dialektalak eta diakronikoak ezagutzen ez dutenei eta entrenamendu corpusetan egitura horiek ez dituzten HPko sistemeni zuzentzen zaie; bigarrena, sinplifikazio naturala, erdi-maila duten ikasleei eta esaldi laburrak hobeto prozesatzen dituzten HPko aplikazioei, eta hirugarrena, sinplifikazio absolutua, euskara ikasten hasi berri direnei eta esaldi bakoitzeko esaldi bakarra prozesatzen duten tresnei.

Kasu azterketa moduan, **Biografix tresna eleaniztuna inplementatu** dugu, EuTS sistemaren eragiketak jarraituz azterketa linguistikoan oinarritutako erregelen bidez egitura parentetikoak sinplifikatzen dituena. Tresna horrekin gure erregelak eta eragiketak sinplifikazio sintaktikoa egiteko baliagarriak direla ikusi dugu. Horretaz gain, euskararako definitutako erregelak beste hizkuntza batzuetan aplika daitezkeela probatu dugu.

Ikerketa-galdera multzo hori 4. eta 6. kapituluetan erantzun dugu.

8.2.3 Baliabideak

TSAko gure hurbilpena kontrastatzeko, **ETSC corpusa osatu** dugu. Bertan jatorrizko hiru testuren eskuz sinplifikatutako bi bertsio bildu ditugu. Horiek analizatzeko, **etiketatze-eskema** bat garatu dugu eta etiketatze-eskema horretan oinarrituta, hurbilpen ezberdinak jarraituta testuak eskuz sinplifikatzean egin diren eragiketak aztertu ditugu. Azterketa hori bi hurbilpenak konparatzeko eta amankomunean dituzten eragiketak lortzeko baliatu dugu. Eragiketa horiekin ETSC corpusa zabaltzeko **bete beharreko eragiketen zerrenda** osatu dugu.

Oinarrizko tresnetan **Mugak eta Aposizioak sortu** ditugu. Testuak sinplifikatu ahal izateko, testuen analisia egingo duten oinarrizko tresnak beharrezkoak dira, eta guk, bereziki, perpausen, aposizioen eta egitura parentetikoen mugak ezartzen dituzten tresnak behar izan ditugu, sinplifikazio sintaktikoan horiek non banatu behar diren jakin behar dugulako. Oinarrizko tresnetan egin ditugun ekarpenak MuGa gramatikaren hobekuntza eta aposizio-detektatzailearen garapena dira. Bi horiek ebaluatzeko urrepatroiak sortu ditugu. Horretaz gain, azterketa linguistikoa egiteko eta sistemak entrenatzeko eta ebaluatzeko **corpusak eta datu-multzoak sortu** ditugu: ErreXail entrenatzeko bi corpus bildu ditugu, *Elhuyar (T-comp)* corpusa eta *Zernola (T-simp)* corpusa. EuTS eta Biografix garatzeko eta

ebaluatzeko bi datu-multzo sortu ditugu: *Wikipedia DS* eta *EGLU DS.*

Ikerketa-galdera multzo hori, batez ere, 4. eta 6. kapituluetan erantzun dugu.

8.2.4 Beste hizkuntzekiko konparazioa

Beste hizkuntzetan egindako lanak aztertu ondoren, euskarak baliabide urriko beste hizkuntzek dituzten behar berak ditu: corpusak eta oinarrizko tresnak. Sinplifikazio-erregeletan euskarak duen ezberdintasuna da erregela horiek ezaugarri morfologikoetan oinarrituta daudela eta hori erronka handia izan daitekeela arkitekturan integratutako moduluentzat. Dena dela, egitura parentetikoen sinplifikazioko kasu-azterketan ikusi dugun bezala, euskararako diseinatu ditugun erregelak beste hizkuntzetan aplika daitezke. Eskuz sinplifikatutako testuen analisian ere ikusi dugu euskaraz eta beste hizkuntzetan egiten diren eragiketak antzekoak direla.

Ikerketa-galdera multzo hori kapitulu guztietan atal bakoitzaren ondoren erantzun dugu.

Ekarpen horiek guztiak 8.1 irudian urdinez nabarmendu ditugu. Irudi horretan gehitu ditugu ekarpen nagusiak bi ikerketa-lerrotan (testuen konplexutasunaren analisia eta testuen konplexutasun sintaktikoaren tratamendua edo sinplifikazioa) banatuta. Oinarrizko tresnetan egindako ekarpenak bi laukizuzenei lotu ditugu.

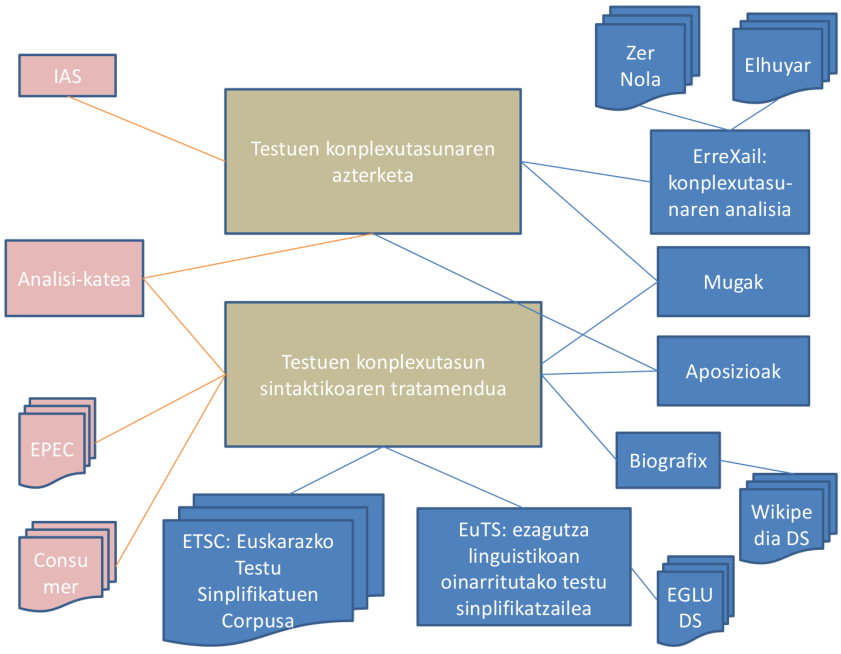
Beraz, tesi-txosten honetan zehar ikusi dugun bezala, ekarpen horietan oinarrituta, **testuak sinplifikatzeko prozesua** hau da:

1. Testuak linguistikoki analizatuko ditugu Ixa taldearen analisi-kateko tresnen bitartez eta garatu ditugun Mugak eta Aposizioak oinarrizko tresnekin.

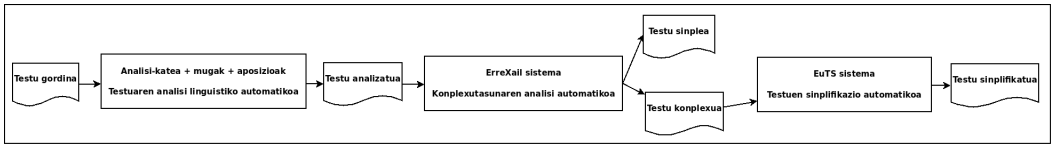
2. Testuen konplexutasuna analizatuko dugu ErreXail sistemarekin.

3. Testuak konplexuak izanez gero, EuTS sistemarekin sinplifikatuko ditugu.

Testuak sinplifikatzeko prozesuaren laburpena 8.2 irudian jaso dugu.



8.1 irudia – Tesian erabilitako baliabideak eta tresnak eta egindako ekarpenak



8.2 irudia – Testuak sinplifikatzeko prozesuaren laburpena

Txosten honen sarrera-kapituluan, 1.1 taulan, aurkeztu dugun esaldira itzuliko gara eta 8.1 taulan jatorrizko esaldiaren itzulpen automatikoarekin batera sinplifikatutako esaldien itzulpen automatikoa[1] emango dugu egindako ekarpenetan oinarrituta.

Jatorrizko esaldiaren itzulpen automatikoan ikusi dugunez, perpausen elementuak nahastu dira eta aditza ez da ondo itzuli. Sinplifikatutako esaldietan, berriz, "elkartu" aditzaren ordainak dira *attended* aditza *the event* eta

[1]Esaldien itzulpen automatikoa egiteko Google Translate `https://translate.google.es/` web zerbitzua erabili dugu 2013ko abenduan.

Jatorrizko esaldia	**Jatorrizko esaldiaren itzulpena**
1962an Charles De Gaulle eta Konrad Adenauer Bonnen elkartu zirenean 55 miloi lagun bizi ziren herrialde horretan, eta 47 milioi Frantzian.	Charles De Gaulle and Konrad Adenauer in Bonn, when 55 million people were living together in this country, and 47 million in France.
Sinplifikatutako esaldiak	**Sinplifikatutako esaldien itzulpena**
1962an Charles De Gaulle eta Konrad Adenauer Bonnen elkartu ziren. Orduan 55 miloi lagun bizi ziren herrialde horretan, eta 47 milioi Frantzian.	Charles De Gaulle and Konrad Adenauer in Bonn in 1962, attended the event. Then, 55 million people were living in the country, and 47 million in France.

8.1 taula – Jatorrizko esaldi baten eta sinplifikatutako esaldien itzulpenak

objektua. Ez da itzulpen zehatza, baina ulergarria da eta ez da jatorrizkotik asko aldatzen. Esaldiko osagaien hurrenkera zuzena ez den arren eta puntuazio-akats bat dagoen arren, esaldiak aditza dauka (ez jatorrizkoaren itzulpenean bezala) eta onargarria[2] da. Sinplifikatutako bigarren esaldiaren itzulpena gramatikalki zuzena da, baina itzulpena guztiz zuzena izateko *that* determinatzailea beharko luke *the* determinatzailea beharrean. Beraz, goiko 8.1 taulako esaldiarekin ikusi dugun bezala, esaldiak itzuli aurretik sinplifikatuta, itzulpenaren emaitza hobetu dugu gramatikaltasunaren eta ulermenaren aldetik.

8.3 Zabaldutako ikerketa-lerroak eta etorkizuneko lanak

Tesi-lan honetan zabaldu ditugun ikerketa-lerroak are gehiago ustia daitezkeela aurreikusten dugu.

- **Testuen konplexutasunaren analisia**
  - Ezaugarriak gehitu: ErreXail sistemari ezaugarri gehiago gehi dakizkioke, analisi semantikoarenak edo hitzen maiztasunak adibidez.

[2]Itzulpen automatikoan, eskuzko ebaluazioak egitean esaldia ulertzen bada, alegia, mezua helarazten bada, onargarria dela esaten da, nahiz eta esaldia guztiz zuzena ez den.

  - Konplexutasun-maila gehiago sailkatu: entrenatzeko testu gehiago lortuz gero, esaterako Hizkuntzetarako Europako Erreferentzia Esparru Bateratuan dauden mailak sailka daitezke.
  - Beste domeinuetara egokitu: esaterako, *Vikidia*ko eta *Wikipedia*ko testuak erabiliz domeinu entziklopedikora egoki daiteke.
  - Testuen analisi estilistikoa: monitorizazio linguistikoari beste erabilera bat eman dakioke testuen ezaugarriak aztertzeko.

- **Testuen sinplifikazio automatikoa**

  - Inplementazioa eta ebaluazioa egin: testuen sinplifikazio automatikoan EuTS sistemaren inplementazioa amaitu eta ebaluatu beharko dugu.

    Ebaluazio intrintsekoa eta estrintsekoa egiteaz gain, esperimentu kognitiboak egin daitezke erabiltzaileekin eta helburu-taldeekin.

    Ebaluazioak oinarrizko tresnen hutsuneak ere erakuts diezazkiguke. Oinarrizko tresnetan dauden gramatiken hobekuntza ere etengabeko lana da; beraz, horien garapenak ere emaitza hobeak lortzea ekarriko luke.
  - Azterketa linguistikoa osatu: inplementazioak eta ebaluazioak ziurrenik erregelen fintzea ekarriko du. Beste domeinuetako corpusetan azterketak eginez, erregelak domeinuetara ere egoki daitezke, eta erregelak jatorri geografikoaren arabera ere molda daitezke.

    Txostenean zehar aipatu dugu egitura parentetiko gehiago analiza ditzakegula eta, ETSC corpuseko emaitzetan ikusi dugun bezala, postposizioak dituzten egituren sinplifikazioa emankorra izan daiteke.

    Corpus handiagoak lortuz gero, perpaus adberbialen corpus azterketa kuantitatiboa zabal daiteke eta esaldien barne-hurrenkera ere azter daiteke.
  - Bestelako ezaugarriak gehitu: testuen sinplifikazio automatikoan sinplifikazio lexikala integratzeaz gain, testuak aberats daitezke estekak gehituz *Wikipedia*ra, mapetara edo webguneetara edo testua definizioekin hornituz.

    Horretaz gain, analisi sintaktikoan oinarritu beharrean, analisi semantikoan oinarritutako probak egin daitezke.

Definitu ditugun hiru sinplifikazio-mailez gain, sinplifikatu nahi den fenomenoa hautatzeko aukera emateko parametriza daiteke. Aukera hori helburu-taldearen beharren araberakoa izango da, eta proposaturiko sinplifikazio-erregela guztien artean hautatuko du testua sinplifikatu nahi duenak.

EuTS sistemaren M-Xuxen moduluan, zuzenketa eta egokitzapena eragiketa egiteko, gramatika-zuzentzailea (XuxenG) eta korreferentziaren ebazpena (Soraluze *et al.*, 2015) integra daitezke.

- **Eskuz sinplifikatutako testuen analisia**
  - ETSC corpusa analizatzen jarrai dezakegu, informazio-ezabatzeak edo zein diren hurrenkera-aldaketak esaterako, eta corpusa zabaltzeari ekin diezaiokegu.
  - Analisi hori sisteman lehentasunak ezartzeko balia dezakegu.

Tesi-lan honetan euskarazko egitura konplexuen analisia egin dugu eta egitura horiek automatikoki sinplifikatzeko bidea ireki dugu.

Bibliografia

Abney S.P. Parsing by Chunks. In Berwick R.C., Abney S.P., eta Tenny C., editors, *Principle-Based Parsing: Computation and Psycholinguistics*. Kluwer Academic, 1991.

Action P.L. eta Network I. *Federal Plain Language Guidelines*. PlainLanguage.gov, 2011.

Aduriz I. *EUSMG: Morfologiatik sintaxira murriztapen gramatika erabiliz [EUSMG: From Morphology to Syntax using Constraint Grammar]*. Doktoretza-tesia, University of the Basque Country (UPV/EHU), 2000.

Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Oronoz M., Sarasola K., Soroa A., Urizar R., eta Urkia M. A framework for the automatic processing of Basque. *Proceedings of Workshop on Lexical Resources for Minority Languages. First LREC Conference. Granada*, 1998.

Aduriz I., Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J.M., Artola X., Gojenola K., Maritxalar M., Sarasola K., eta Urkia M. A Word-grammar Based Morphological Analyzer for Agglutinative Languages. *COLING, 18th International Conference on Computational Linguistics*. Universitaet des Saarlandes, Saarbrueucken, Germany, Morgan Kaufmann, 2000.

Aduriz I., Aranzabe M.J., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., eta Urizar R. Methodology and Steps Towards the Construction of EPEC, a Corpus of Written Basque Tagged at Morphological and Syntactic levels for Automatic Processing. *Language and Computers*, 56(1):1–15, 2006a.

Aduriz I., Aranzabe M.J., Arriola J.M., Díaz de Ilarraza A., Gojenola K., Oronoz M., eta Uria L. A Cascaded Syntactic Analyser for Basque. *Computational Linguistics and Intelligent Text Processing*, 124–134, 2004.

Aduriz I., Arrieta B., Arriola J.M., Díaz de Ilarraza A., Izagirre E., eta Ondarra A. Muga Gramatikaren Optimizazioa [Optimization of the Clause Boundary Grammar]. Barne-txostena, UPV/EHU/LSI/TR 9-2006, 2006b.

Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., eta Maritxalar M. Morphosyntactic Disambiguation for Basque Based on the Constraint Grammar Formalism. *Proceedings of Recent Advances on NLP (RANLP)*, 282–288, Tzigov Chark, Bulgary, 1997.

Aduriz I., Arriola J.M., Gonzalez-Dios I., eta Urizar R. Funtzio Sintaktikoen Gold Estandarra eskuz etiketatzeko gidalerroak [Guidelines to Annotate the Gold-standard of Syntactic Functions]. Barne-txostena, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 01-2015, 2015.

Aduriz I. eta Díaz de Ilarraza A. Morphosyntactic Disambiguation and Shallow Parsing in Computational Processing of Basque. *Inquiries into the lexicon-syntax relations in Basque*. Servicio Editorial de la Universidad del Pais Vasco-Euskal Herriko Unibertsirareko Argitalpen Zerbitzua, 2003.

Agirre E., Aldezabal I., Etxeberria J., Iruskieta M., Izagirre E., Mendizabal K., eta Pociello E. A Methodology for the Joint Development of the Basque WordNet and Semcor. *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC). Genoa (Italy)*, 2006.

Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Maritxalar M., Sarasola K., eta Urkia M. Xuxen: A Spelling Checker/Corrector for Basque based in Two-Level Morphology. *Proceedings of NAACL-ANLP'92, 119-125. Povo Trento*, 1992.

Agirrezabal M., Gonzalez-Dios I., eta Lopez-Gazpio I. Euskararen Sorkuntza Automatikoa: lehen urratsak [Automatic Generation of Basque: First

Steps]. *I. Ikergazte Nazioarteko ikerketa euskaraz Kongresuko artikulubilduma*, 15–23, 2015.

Aha D.W., Kibler D., eta Albert M.C. Instance-based Learning Algorithms. *Machine Learning*, 6:37–66, 1991.

Al-Ajlan A.A., Al-Khalifa H.S., eta Al-Salman A. Towards the Development of an Automatic Readability Measurements for Arabic Language. *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, 506–511. IEEE, 2008.

Al-Subaihin A.A. eta Al-Khalifa H.S. Al-Baseet: A proposed Simplification Authoring Tool for the Arabic Language. *International Conference on Communications and Information Technology (ICCIT)*, 121–125, 2011.

Al Tarouti F., Kalita J., eta McGrory C. Sentence Simplification for Question Generation. *International Conference on Computing and Communication Systems*, 2015.

Alcázar A. Towards Linguistically Searchable Text. *Proceedings of BIDE Summer School of Linguistics*, 2005.

Aldabe I., Gonzalez-Dios I., Lopez-Gazpio I., Madrazo I., eta Maritxalar M. Two Approaches to Generate Questions in Basque. *Procesamiento del Lenguaje Natural*, 51:101–108, 2013.

Aldabe I., Maritxalar M., Perez de Viñaspre O., eta Larraitz U. Automatic Exercise Generation in an Essay Scoring System. *Proceedings of the 20th International Conference on Computers in Education*, 671–673, 2012.

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., eta Lersundi M. EDBL: a General Lexical Basis for the Automatic Processing of Basque. *Proceedings of the IRCS Workshop on linguistic databases*, 2001.

Aldezabal I., Aranzabe M.J., Arriola J.M., Díaz de Ilarraza A., Estarrona A., Fernandez K., Iruskieta M., eta Uria L. EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) dependentziekin etiketatzeko eskuliburua [Guidelines to Annotate EPEC (the Reference Corpus for the Processing of Basque) with Dependencies]. Barne-txostena, UPV/EHU / LSI / TR 12-2007, 2007a.

Aldezabal I., Aranzabe M.J., Díaz de Ilarraza A., Estarrona A., Fernandez K., eta Uria L. EPEC-RS: EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) rol semantikoekin etiketatzeko eskuliburua [Guidelines to Annotate EPEC-RS (the Reference Corpus for the Processing of Basque) with Semantic Roles]. Barne-txostena, UPV/EHU/LSI/TR 02-2010, 2010.

Aldezabal I., Ceberio K., Esparza I., Estarrona A., Etxeberria J., Iruskieta M., Izagirre E., eta Uria L. EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) segmentazio-mailan etiketatzeko eskuliburua [Guidelines to Annotate EPEC (the Reference Corpus for the Processing of Basque) at Segmentation Level]. Barne-txostena, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 11-2007, 2007b.

Alegria I. *Euskal morfologiaren tratamendu automatikorako tresnak [Tools for the Treatment of Basque Morphology]*. Doktoretza-tesia, University of the Basque Country (UPV/EHU), 1995.

Alegria I., Ansa O., Artola X., Ezeiza N., Gojenola K., eta Urizar R. Representation and Treatment of Multiword Expressions in Basque. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, 48–55. Association for Computational Linguistics, 2004.

Alegria I., Ezeiza N., Fernandez I., eta Urizar R. Named Entity Recognition and Classification for Texts in Basque. *II Jornadas de Tratamiento y Recuperación de Información, JOTRI, Madrid*, 2003. ISBN 84-89315-33-7.

Allen D. A study of the Role of Relative Clauses in the Simplification of News Texts for Learners of English. *System*, 37(4):585–599, 2009.

Aluísio S., Specia L., Gasperin C., eta Scarton C. Readability Assessment for Text Simplification. *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, 1–9. Association for Computational Linguistics, 2010.

Aluísio S.M. eta Gasperin C. Fostering Digital Inclusion and Accessibility: The PorSimples Project for Simplification of Portuguese Texts. *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, 46–53. Association for Computational Linguistics, 2010.

Aluísio S.M., Specia L., Pardo T.A.S., Maziero E.G., Caseli H.M., eta Fortes R.P.M. A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps Towards Text Simplification Systems. *Proceedings of the 26th annual ACM international conference on Design of communication*, SIGDOC '08, 15–22, New York, NY, USA, 2008a. ACM. ISBN 978-1-60558-083-8. URL http://doi.acm.org/10.1145/1456536.1456540.

Aluísio S.M., Specia L., Pardo T.A., Maziero E.G., eta Fortes R.P. Towards Brazilian Portuguese Automatic Text Simplification Systems. *Proceedings of the eighth ACM symposium on Document engineering*, DocEng '08, 240–248, New York, NY, USA, 2008b. ACM. ISBN 978-1-60558-081-4. URL http://doi.acm.org/10.1145/1410140.1410191.

Amoia M. eta Romanelli M. SB: mmSystem-Using Decompositional Semantics for Lexical Simplification. *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 482–486. Association for Computational Linguistics, 2012.

Angrosh M. eta Siddharthan A. Text Simplification Using Synchronous Dependency Grammars: Generalising Automatically Harvested Rules. *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, 16–25, Philadelphia, Pennsylvania, U.S.A., June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-4404.

Aranberri N., Labaka G., Díaz de Ilarraza A., eta Sarasola K. Comparison of Post-editing Productivity between Professional Translators and Lay Users. *Proceedings of Third Workshop on Post-Editing Technology and Practice*, 20–33, 2014.

Aranzabe M.J. *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala [Syntactic Resources based on the Dependency Model: the Treebank and the Computational Grammar]*. Doktoretza-tesia, University of the Basque Country (UPV/EHU), 2008.

Aranzabe M.J., Díaz de Ilarraza A., eta Gonzalez-Dios I. First Approach to Automatic Text Simplification in Basque. In Rello L. eta Saggion H., edi-

tors, *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, 1–8, 2012a.

Aranzabe M.J., Díaz de Ilarraza A., eta Gonzalez-Dios I. Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento de Lenguaje Natural*, 50:61–68, 2013.

Aranzabe M.J., Kepa B., de Ilarraza Arantza D., Nerea E., Goenaga I., eta Gojenola K. Combining Rule-Based and Statistical Syntactic Analyzers. *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012), pp. 48-54, Association for Computational Linguistics (ACL), USA, ISBN: 978-1-937284-30-5, July 12, 2012, Jeju Island, Republic of Korea*, 2012b.

Arrieta B. *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean [The Treatment of the Surface Syntax with Machine Learning Techniques: the Identification of Basque Chunks and their Use in a Comma Checker]*. Doktoretza-tesia, University of the Basque Country (UPV/EHU), 2010.

Bach N., Gao Q., Vogel S., eta Waibel A. TriS: A Statistical Sentence Simplifier with Log-linear Models and Margin-based Discriminative Training. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 474–482, 2011.

Baeza-Yates R., Rello L., eta Dembowski J. CASSA: A Context-Aware Synonym Simplification Algorithm. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1380–1385, Denver, Colorado, May–June 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N15-1156.

Barbu E., Martín-Valdivia M.T., Martínez-Cámara E., eta Ureña-López L.A. Language Technologies Applied to Document Simplification for Helping Autistic People. *Expert Systems with Applications*, 42(12):5076–5086, 2015.

Barlacchi G. eta Tonelli S. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. *Computational Linguistics and Intelligent Text Processing*, 476–487. Springer, 2013.

Bautista S., Drndarevic B., Hervás R., Saggion H., eta Gervás P. Análisis de la Simplificación de Expresiones Numéricas en Español mediante un Estudio Empírico. *Linguamática*, 4(2):27–41, 2012a.

Bautista S., Hervás R., eta Gervás P. Simplificación de textos centrada en la adaptación de expresiones numéricas. *I Congreso Internacional Universidad y Discapacidad*, Madrid, 2012b.

Bautista S., Hervás R., Gervás P., Power R., eta Williams S. A System for the Simplification of Numerical Expressions at Different Levels of Understandability. *Natural Language Processing for Improving Textual Accessibility (NLP4ITA 2013)*, 10–19, 2013.

Bautista S., Hervás R., Gervás P., eta Rojo J. An Approach to Treat Numerical Information in the Text Simplification Process. *Universal Access in the Information Society*, 1–18, 2015.

Bautista S. eta Saggion H. Making Numerical Information more Accessible. The Implementation of a Numerical Expression Simplification System for Spanish. In François T. eta Bernhard D., editors, *International Journal of Applied Linguistics. Special Issue on Recent Advances in Automatic Readability Assessment and Text Simplification*, 165 lib., 299–323. John Benjamins Publishing Company, 2014.

Bawakid A. eta Oussalah M. Sentences Simplification for Automatic summarization. *IEEE 10th International Conference on Cybernetic Intelligent Systems (CIS)*, 59–64. IEEE, 2011.

Beigman Klebanov B., Knight K., eta Marcu D. Text Simplification for Information-Seeking Applications. *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, 735–747, 2004.

Bengoetxea K. *Estaldura zabaleko euskararako analizatzaile sintaktiko estatistikoa [The Statistical Parser of Basque with Extensive Coverage]*. Doktoretza-tesia, University of the Basque Country (UPV/EHU), 2014.

Bengoetxea K. eta Gojenola K. Desarrollo de un analizador sintáctico estadístico basado en dependencias para el euskera. *Procesamiento del lenguaje natural*, 39:5–12, 2007.

Benjamin R.G. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educational Psychology Review*, 24(1):63–88, 2012.

Bernhard D., De Viron L., Moriceau V., eta Tannier X. Question Generation for French: Collating Parsers and Paraphrasing Questions. *Dialogue and Discourse*, 3(2):43–74, 2012.

Biran O., Brody S., eta Elhadad N. Putting it Simply: a Context-Aware Approach to Lexical Simplification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 496–501, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-2087.

Blake C., Kampov J., Orphanides A.K., West D., eta Lown C. UNC-CH at DUC 2007: Query Expansion, Lexical Simplification and Sentence Selection Strategies for Multi-Document Summarization. *Proceedings of the Document Understanding Conference*, 2007.

Bott S., Rello L., Drndarevic B., eta Saggion H. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. *Proceedings of COLING*, 357–373, 2012a.

Bott S. eta Saggion H. An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, 20–26, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 9781937284053. URL http://dl.acm.org/citation.cfm?id=2107679.2107682.

Bott S. eta Saggion H. Automatic Simplification of Spanish Text for E-accessibility. *Computers Helping People with Special Needs*, 527–534. Springer, 2012.

Bott S. eta Saggion H. Text Simplification Resources for Spanish. *Language Resources and Evaluation*, 48(1):93–120, 2014.

Bott S., Saggion H., eta Figueroa D. A Hybrid System for Spanish Text Simplification. *Third Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 75–84, Montreal, Canada, 2012b.

Bott S., Saggion H., eta Mille S. Text Simplification Tools for Spanish. In Calzolari (Conference Chair) N., Choukri K., Declerck T., Uğur Doğan M., Maegaard B., Mariani J., Odijk J., eta Piperidis S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 1665–1671, Istanbul, Turkey, May 2012c. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Breiman L. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

Brouwers L., Bernhard D., Ligozat A.L., eta François T. Simplification syntaxique de phrases pour le français. *Actes de la Conférence Conjointe JEP-TALN-RECITAL, Montpellier, France*, 211–224, 2012.

Brouwers L., Bernhard D., Ligozat A.L., eta Francois T. Syntactic Sentence Simplification for French. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 47–56, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-1206.

Brunato D. *A Study on Linguistic Complexity from a Computational Linguistics Perspective. A Corpus-based Investigation of Italian Bureaucratic Texts*. Doktoretza-tesia, Universitá di Siena, 2015.

Brunato D., Dell'Orletta F., Venturi G., eta Montemagni S. Design and Annotation of the First Italian Corpus for Text Simplification. *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, 31–41, 2015.

Burstein J. Opportunities for Natural Language Processing Research in Education. In Gelbukh A., editor, *Computational Linguistics and Intelligent Text Processing*, 5449 lib. of *Lecture Notes in Computer Science*, 6–27. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-00381-3. URL http://dx.doi.org/10.1007/978-3-642-00382-0_2.

Buyko E., Faessler E., Wermter J., eta Hahn U. Syntactic Simplification and Semantic Enrichment-Trimming Dependency Graphs for Event Extraction. *Computational Intelligence*, 27(4):610–644, 2011.

Candido A. Jr., Maziero E., Gasperin C., Pardo T.A.S., Specia L., eta Aluísio S.M. Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese. *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, 34–42, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-37-4. URL http://dl.acm.org/citation.cfm?id=1609843.1609848.

Canning Y. eta Tait J. Syntactic Simplification of Newspaper Text for Aphasic Readers. *ACM SIGIR'99 Workshop on Customised Information Delivery*, 6–11. Citeseer, 1999.

Carroll J., Minnen G., Canning Y., Devlin S., eta Tait J. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, 7–10. Citeseer, 1998.

Carroll J., Minnen G., Pearce D., Canning Y., Devlin S., eta Tait J. Simplifying Text for Language-impaired Readers. *Proceedings of EACL*, 99 lib., 269–270. Citeseer, 1999.

Caseli H.M., Pereira T.F., Specia L., Pardo T.A.S., Gasperin C., eta Aluísio S. Building a Brazilian Portuguese Parallel Corpus of Original and Simplified Texts. *the Proceedings of CICLing*, 59–70, 2009.

Castro-Castro D., Lannes-Losada R., Maritxalar M., Niebla I., Pérez-Marqués C., Alamo-Suarez N.C., eta Pons-Porrata A. A Multilingual Application for Automated Essay Scoring. *Lecture Notes in Advances in Artificial Intelligence - LNAI 5290 - IBERAMIA*, 243–251. Springer New York, 2008. ISBN 3-540-99308-8.

Chall J.S. eta Dale E. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge, MA, 1995.

Chandrasekar R., Doran C., eta Srinivas B. Motivations and Methods for Text Simplification. *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, 1041–1044, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/993268.993361.

Chandrasekar R. eta Srinivas B. Automatic Induction of Rules for Text Simplification. *Knowledge-Based Systems*, 10(3):183–190, 1997.

Chang C.C. eta Lin C.J. LIBSVM - A Library for Support Vector Machines, 2001. URL http://www.csie.ntu.edu.tw/~cjlin/libsvm/. The Weka classifier works with version 2.82 of LIBSVM.

Chen H.B., Huang H.H., Chen H.H., eta Tan C.T. A Simplification-Translation-Restoration Framework for Cross-Domain SMT Applications. *COLING*, 545–560, 2012.

Chen Y.H., Tsai Y.H., eta Chen Y.T. Chinese Readability Assessment Using TF-IDF and SVM. *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, 2 lib., 705–710. IEEE, 2011.

Chung J.W., Min H.J., Kim J., eta Park J.C. Enhancing Readability of Web Documents by Text Augmentation for Deaf People. *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, WIMS '13, 30:1–30:10, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1850-1. URL http://doi.acm.org/10.1145/2479787.2479808.

Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

Collados J.C. Splitting Complex Sentences for Natural Language Processing Applications: Building a Simplified Spanish Corpus. *Procedia-Social and Behavioral Sciences*, 95:464–472, 2013.

Collantes M., Hipe M., Sorilla J.L., Tolentino L., eta Samson B. Simpatico: A Text Simplification System for Senate and House Bills. *The 11th National Natural Language Processing Research Symposium*, 2015.

Coster W. eta Kauchak D. Learning to Simplify Sentences Using Wikipedia. *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, 1–9, Stroudsburg, PA, USA, 2011a. Association for Computational Linguistics. ISBN 9781937284053. URL http://dl.acm.org/citation.cfm?id=2107679.2107680.

Coster W. eta Kauchak D. Simple English Wikipedia: A New Text Simplification Task. *ACL (Short Papers)'11*, 665–669, 2011b.

Crossley S.A., Allen D., eta McNamara D.S. Text Simplification and Comprehensible Input: A case for an Intuitive Approach. *Language Teaching Research*, 16(1):89–108, 2012.

Daelemans W., Höthker A., eta Sang E.T.K. Automatic Sentence Simplification for Subtitling in Dutch and English. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 1045–1048, 2004.

Damay J.J.S., Lojico G.J.D., Lu K.A.L., Tarantan D.B., eta Ong E.C. SIMTEXT. Text Simplification of Medical Literature. *3rd National Natural Language Processing Symposium - Building Language Tools and Resources*, 34–38, 2006.

Daud N.M., Hassan H., eta Aziz N.A. A Corpus-Based Readability Formula for Estimate of Arabic Texts Reading Difficulty. *World Applied Sciences Journal*, 21:168–173, 2013.

De Belder J., Deschacht K., eta Moens M.F. Lexical Simplification. *Proceedings of Itec2010: 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*, 2010. URL https://lirias.kuleuven.be/handle/123456789/268437.

De Belder J. eta Moens M.F. Text Simplification for Children. *Proceedings of the SIGIR workshop on accessible search systems*, 19–26, 2010.

De Belder J. eta Moens M.F. A Dataset for the Evaluation of Lexical Simplification. *Computational Linguistics and Intelligent Text Processing*, 426–437, 2012.

Dehé N. eta Kavalova Y. Parentheticals. An introduction. In Dehé N. eta Kavalova Y., editors, *Parentheticals*, 1–22. John Benjamins Publishing Company, 2007.

Dell'Orletta F., Montemagni S., eta Venturi G. READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, SLPAT '11, 73–83, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-14-5. URL http://dl.acm.org/citation.cfm?id=2140499.2140511.

Devlin S. eta Unthank G. Helping Aphasic People Process Online Information. *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, Assets '06, 225–226, New York, NY, USA, 2006. ACM. ISBN 1-59593-290-9. URL http://doi.acm.org/10.1145/1168987.1169027.

Dingli A. eta Cachia C. Adaptive eBook. *Interactive Mobile Communication Technologies and Learning (IMCL), 2014 International Conference on*, 14–19. IEEE, 2014.

Doi T. eta Sumita E. Splitting Input Sentence for Machine Translation Using Language Model with Sentence Similarity. *Proc. of the 20th international conference on Computational Linguistics*, 2004.

Drndarević B., Štajner S., Bott S., Bautista S., eta Saggion H. Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. *Computational Linguistics and Intelligent Text Processing*, 488–500. Springer, 2013.

DuBay W.H. The Principles of Readability. *Impact Information*, 1–76, 2004.

Erdocia K., Laka I., Mestres-Missé A., eta Rodriguez-Fornells A. Syntactic Complexity and Ambiguity Resolution in a Free Word Order Language: Behavioral and Electrophysiological Evidences from Basque. *Brain and language*, 109(1):1–17, 2009.

Euskaltzaindia. V, (Mendeko perpausak-1, osagarriak, erlatiboak, konparaziozkoak, ondoriozkoak) [V (Subordinate Clauses-1, Completive, Relative, Comparative, Consecutive]. *Euskal Gramatika Lehen Urratsak [Basque Grammar First Steps]*. Euskaltzaindia, Bilbo, 1999.

Euskaltzaindia. *Euskal Gramatika Laburra: perpaus bakuna*. Euskaltzaindia, Bilbo, 2002.

Euskaltzaindia. VI, (Mendeko perpausak-2, baldintzazkoak, denborazkoak, helburuzkoak, kausazkoak, kontseziozkoak eta moduzkoak) [VI (Subordinate Clauses-2, Conditional, temporal, Purpose, Causal, Concessive and Modal]. *Euskal Gramatika Lehen Urratsak [Basque Grammar First Steps]*. Euskaltzaindia, Bilbo, 2005.

Euskaltzaindia. VII, (Perpaus jokatugabeak: denborazkoak, kausazkoak eta helburuzkoak, baldintzazkoak, kontzesiozkoak, moduzkoak, erlatiboak eta osagarriak) [VII (Subordinate Clauses-2, temporal, Causal and Purpose, Conditional, Concessive, Modal, Relative and Completive]. *Euskal Gramatika Lehen Urratsak [Basque Grammar First Steps]*. Euskaltzaindia, Bilbo, 2011.

Evans R., Orasan C., eta Dornescu I. An Evaluation of Syntactic Simplification Rules for People with Autism. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 131–140, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-1215.

Evans R.J. Comparing Methods for the Syntactic Simplification of Sentences in Information Extraction. *Literary and linguistic computing*, 26(4):371–388, 2011.

Ezeiza N. *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile morfosintaktiko sendo eta malgua.* Doktoretza-tesia, University of the Basque Country (UPV/EHU), 2002.

Fajardo I., Tavares G., Ávila V., eta Ferrer A. Towards text simplification for poor readers with intellectual disability: When do connectives enhance text cohesion? *Research in Developmental Disabilities*, 34(4):1267–1279, 2013.

Falkenjack J., Mühlenbock K.H., eta Jönsson A. Features indicating readability in Swedish text. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, 27–40, 2013.

Feblowitz D. eta Kauchak D. Sentence Simplification as Tree Transduction. *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, 1–10, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2901.

Fellbaum C. WordNet. In Poli R., Healy M., eta Kameas A., editors, *Theory and Applications of Ontology: Computer Applications*, 231–243. Springer

Netherlands, 2010. ISBN 978-90-481-8846-8. URL http://dx.doi.org/10.1007/978-90-481-8847-5_10.

Feng L. Automatic Readability Assessment for People with Intellectual Disabilities. *SIGACCESS Access. Comput.*, (93):84–91, January 2009. ISSN 1558-2337. URL http://doi.acm.org/10.1145/1531930.1531940.

Feng L., Jansche M., Huenerfauth M., eta Elhadad N. A Comparison of Features for Automatic Readability Assessment. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 276–284. Association for Computational Linguistics, 2010.

Fernandez Gonzalez I. *Euskarazko Entitate-Izenak: identifikazioa, sailkapena, itzulpena eta desanbiguazioa.* Doktoretza-tesia, University of the Basque Country (UPV/EHU), 2012.

Ferrés D., Marimon M., eta Saggion H. A Web-based Text Simplification System for English. *Procesamiento del Lenguaje Natural*, 55:191–194, 2015.

Finegan-Dollak C. eta Radev D.R. Sentence Simplification, Compression, and Disaggregation for Summarization of Sophisticated Documents. *Journal of the Association for Information Science and Technology*, 2015.

Flesch R. A new readability yardstick. *Journal of applied psychology*, 32(3): 221, 1948.

François T. eta Fairon C. An AI Readability Formula for French as a Foreign Language. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 466–477. Association for Computational Linguistics, 2012.

Freyhoff G., Hess G., Kerr L., Menzel E., Tronbacke B., eta Van Der Veken K. *Make it Simple. European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability.* ILSMH European Association, 1998.

Gala N., François T., eta Fairon C. Towards a French Lexicon with Difficulty Measures: NLP Helping to Bridge the Gap between Traditional Dictionaries and Specialized Lexicons. In Kosem I., Kallas J., Gantar P., Krek S., Langemets M., eta Tuulik M., editors, *Electronic lexicography in the 21st*

century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia., 132–151, Ljubljana/Tallinn, 2013. Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Gasperin C., Maziero E., eta Aluisio S.M. Challenging Choices for Text Simplification. *Computational Processing of the Portuguese Language*, 40–50. Springer, 2010.

Gasperin C., Maziero E., Specia L., Pardo T.A., eta Aluisio S.M. Natural Language Processing for Social Inclusion: a Text Simplification Architecture for Different Literacy Levels. *the Proceedings of SEMISH-XXXVI Seminário Integrado de Software e Hardware*, 387–401, 2009a.

Gasperin C., Specia L., Pereira T., eta Aluísio S. Learning when to Simplify Sentences for Natural Text Simplification. *Proceedings of ENIA*, 809–818, 2009b.

Gojenola K. *Euskararen sintaxi konputazionalerantz. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan [Towards the Computational Syntax of Basque. Basic Resources and their Application in the Extraction of Verb Subcategorisation Information and Error Treatment]*. Doktoretza-tesia, University of the Basque Country (UPV/EHU), 2000.

Gonzalez-Dios I. Euskarazko egitura sintaktikoen azterketa testuen sinplifikazio automatikorako: Aposizioak, erlatibozko perpausak eta denborazko perpausak [Analysis of Basque Syntactic Structures for Automatic Text Simplification]. Master-tesia, University of the Basque Country (UPV/EHU), 2011.

Gonzalez-Dios I. Euskarazko testuen sinplifikazio automatikoa [Automatic Simplification of Basque Texts]. Abstract and oral presentation, 2013.

Gonzalez-Dios I. Euskarazko testuak errazten: euskal testuen sinplifikazio automatikoa [Making Basque Texts Easier: Automatic Simplification of Basque Texts]. In Aduriz I. eta Urizar R., editors, *Euskal Hizkuntzalaritzan egungo zenbait ikerlerro. Hizkuntzalari euskaldunen I. topaketa*, 135–149. Udako Euskal Unibertsitatea, 2014a.

Gonzalez-Dios I. Simplificación automática de textos en euskera [Automatic Simplification of Basque Texts]. In Ureña López L.A., Troyano Jiménez

J.A., Ortega Rodríguez F.J., eta Martínez Cámara E., editors, *Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org*, 45–50, 2014b.

Gonzalez-Dios I., Aranzabe M.J., eta de Ilarraza A.D. Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification. *Proceedings the 7th Language & Technology Conference.*, 450–454, 2015a.

Gonzalez-Dios I., Aranzabe M.J., de Ilarraza A.D., eta Soraluze A. Detecting Apposition for Text Simplification in Basque. *Computational Linguistics and Intelligent Text Processing*, 513–524. Springer, 2013a.

Gonzalez-Dios I., Aranzabe M.J., eta Díaz de Ilarraza A. Testuen sinplifikazio automatikoa: arloaren egungo egoera [Automatic Text Simplification: State of Art]. *Linguamática*, 5(2):43–63, Dezembro 2013b. ISSN 1647–0818. URL http://www.linguamatica.com/index.php/linguamatica/article/view/163.

Gonzalez-Dios I., Aranzabe M.J., eta Díaz de Ilarraza A. Making Biographical Data in Wikipedia Readable: A Pattern-based Multilingual Approach. *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, 11–20, Dublin, Ireland, August 2014a. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/W14-5602.

Gonzalez-Dios I., Aranzabe M.J., eta Díaz de Ilarraza A. Perpaus adberbialen agerpena, maiztasuna eta kokapena EPEC-DEP corpusean [Presence, frequency and Position of Basque Adberbial Clauses in The BDP corpus]. Barne-txostena, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 02-2015, 2015b.

Gonzalez-Dios I., Aranzabe M.J., Díaz de Ilarraza A., eta Salaberri H. Simple or Complex? Assessing the Readability of Basque Texts. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 334–344, Dublin, Ireland, August 2014b. Dublin City University and Association for Computational Linguistics. URL http://www.aclweb.org/anthology/C14-1033.

Graesser A.C., McNamara D.S., Louwerse M.M., eta Cai Z. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36 (2):193–202, 2004.

Gunning R. *The Technique of Clear Writing*. McGraw-Hill New York, 1968.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., eta Witten I.H. The WEKA Data Mining Software: an Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

Hancke J., Vajjala S., eta Meurers D. Readability Classification for German using Lexical, Syntactic, and Morphological Features. *COLING 2012: Technical Papers*, 1063–1080, 2012.

Heilman M. eta Smith N.A. Extracting Simplified Statements for Factual Question Generation. *Proceedings of QG2010: The Third Workshop on Ques-tion Generation*, 2010.

Hervás R., Bautista S., Rodríguez M., de Salas T., Vargas A., eta Gervás P. Integration of Lexical and Syntactic Simplification Capabilities in a Text Editor. *Procedia Computer Science*, 27:94–103, 2014.

Hualde J.I. eta Ortiz de Urbina J., editors. *A Grammar of Basque*. Mouton de Gruyter, 2003.

Hulden M. Foma: a Finite-State Compiler and Library. *EACL (Demos)'09*, 29–32, 2009.

Hwang W., Hajishirzi H., Ostendorf M., eta Wu W. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2015.

Inui K., Fujita A., Takahashi T., Iida R., eta Iwakura T. Text Simplification for Reading Assistance: A Project Note. *Proceedings of the second international workshop on Paraphrasing-Volume 16*, 9–16. Association for Computational Linguistics, 2003.

Iruskieta M., Aranzabe M.J., Díaz de Ilarraza A., Gonzalez-Dios I., Lersundi M., eta Lopez de Lacalle O. The RST Basque TreeBank: an Online Search Interface to Check Rhetorical Relations. *Proceedings of the 4th Workshop RST and Discourse Studies*, 40–49, 2013.

Iruskieta M., Díaz de Ilarraza A., eta Lersundi M. Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera. *Procesamiento del Lenguaje Natural*, 47:137–144, 2011.

Ixa Taldea. *Ixa taldeko etiketen eskuliburua*, 2004.

Jauhar S.K. eta Specia L. UOW-SHEF: SimpLex–Lexical Simplicity Ranking based on Contextual and Psycholinguistic Features. *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 477–481. Association for Computational Linguistics, 2012.

Johannsen A., Martínez H., Klerke S., eta Søgaard A. EMNLP@ CPH: Is Frequency all there is to Simplicity? *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 408–412. Association for Computational Linguistics, 2012.

John G.H. eta Langley P. Estimating Continuous Distributions in Bayesian Classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence*, 338–345, San Mateo, 1995. Morgan Kaufmann.

Jonnalagadda S. eta Gonzalez G. BioSimplify: an Open Source Sentence Simplification Engine to Improve Recall in Automatic Biomedical Information Extraction. *AMIA Annual Symposium Proceedings*, 2010 lib., 351–355. American Medical Informatics Association, 2010a.

Jonnalagadda S. eta Gonzalez G. Sentence Simplification Aids Protein-Protein Interaction Extraction. *Arxiv preprint arXiv:1001.4273*, 2010b.

Jonnalagadda S., Tari L., Hakenberg J., Baral C., eta Gonzalez G. Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 177–180. Association for Computational Linguistics, 2009.

Kajiwara T. eta Yamamoto K. Evaluation Dataset and System for Japanese Lexical Simplification. *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, 35–40, Beijing, China, July 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P15-3006.

Kandula S., Curtis D., eta Zeng-Treitler Q. A Semantic and Syntactic Text Simplification Tool for Health Content. *AMIA Annual Symposium Proceedings*, 2010 lib., 366–370. American Medical Informatics Association, 2010.

Karlsson F., Voutilainen A., Heikkila J., eta Anttila A. *Constraint Grammar, A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter, 1995.

Kauchak D. Improving Text Simplification Language Modeling Using Unsimplified Text Data. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1537–1546, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P13-1151.

Kauchak D., Mouradi O., Pentoney C., eta Leroy G. Text Simplification Tools: Using Machine Learning to Discover Features that Identify Difficult Text. *47th Hawaii International Conference on System Sciences (HICSS)*, 2616–2625. IEEE, 2014.

Keskisärkkä R. Automatic Text Simplification via Synonym Replacement. Master-tesia, Linköping, 2012. URL http://liu.diva-portal.org/smash/get/diva2:560901/FULLTEXT01.

Kim Y.S., Hullman J., eta Adar E. DeScipher: A Text Simplification Tool for Science Journalism. *Computation+Journalism Symposium*, 2015.

Klaper D., Ebling S., eta Volk M. Building a German/Simple German Parallel Corpus for Automatic Text Simplification. *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, 11–19, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2902.

Klerke S. eta Søgaard A. DSim, a Danish Parallel Corpus for Text Simplification. In Calzolari (Conference Chair) N., Choukri K., Declerck T., Uğur Doğan M., Maegaard B., Mariani J., Odijk J., eta Piperidis S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 4015–4018, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Klerke S. eta Søgaard A. Simple, Readable Sub-sentences. *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, 142–149, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P13-3021.

Kvistab M. eta Velupillaia S. Professional Language in Swedish Radiology Reports–Characterization for Patient-Adapted Text Simplification. *Scandinavian Conference on Health Informatics 2013*, 55–59, 2013.

Labaka G. *EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation.* Doktoretza-tesia, University of the Basque Country (UPV/EHU), Donostia, March 2010.

Laka I. A Brief Grammar of Euskara, the Basque Language, 1996. URL http://www.ehu.es/grammar.

Lal P. eta Rüger S. Extract-based Summarization with Simplification. *Proceedings of the Workshop on Text Summarization at DUC 2002 In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization*, 2002.

Lasecki W.S., Rello L., eta Bigham J.P. Measuring Text Simplification with the Crowd. *Proceedings of the 12th Web for All Conference*, W4A '15, 4:1–4:9, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3342-9. URL http://doi.acm.org/10.1145/2745555.2746658.

Leroy G., Endicott J.E., Kauchak D., Mouradi O., eta Just M. User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiarity on Perception, Understanding, Learning, and Information Retention. *Journal of medical Internet research*, 15(7), 2013.

Ligozat A.L., Garcia-Fernandez A., Grouin C., eta Bernhard D. ANNLOR: A Naïve Notation-system for Lexical Outputs Ranking. *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 487–492. Association for Computational Linguistics, 2012.

Ligozat A.L., Grouin C., Garcia-Fernandez A., eta Bernhard D. Approches à base de fréquences pour la simplification lexicale. *Actes TALN-RÉCITAL 2013*, 493–506. ATALA, 2013.

Lopez-Gazpio I. eta Maritxalar M. Web application for Reading Practice. In IADAT, editor, *IADAT-e2013: Proceedings of the 6th IADAT International Conference on Education*, 74–78, 2013.

Lozanova S., Stoyanova I., Leseva S., Koeva S., eta Savtchev B. Text Modification for Bulgarian Sign Language Users. *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, 39–48, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2905.

Madrazo I. Testuen irakurgarritasuna neurtzeko sailkatzaile automatikoa [An Automatic Classiffier of Text Legibility]. Master-tesia, University of the Basque Country (UPV/EHU), 2014.

Max A. Simplification interactive pour la production de textes adaptés aux personnes souffrant de troubles de la compréhension. *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, 2005.

Max A. Writing for Language-Impaired Readers. In Gelbukh A., editor, *Computational Linguistics and Intelligent Text Processing*, 3878 lib. of *Lecture Notes in Computer Science*, 567–570. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-32205-4. URL http://dx.doi.org/10.1007/11671299_59.

Medero J. eta Ostendorf M. Identifying Targets for Syntactic Simplification. *Proceedings of the SLaTE 2011 workshop*, 69–72, 2011.

Mencap. *Am I Making Myself Clear? Mencap's Guidelines for Accessible Writing*. Mencap, 2000.

Minard A.L., Ligozat A.L., eta Grau B. Simplification de phrases pour l'extraction de relations. *Proceedings of the Joint Conference JEP-TALN-RÉCITAL 2012, volume 2: TALN*, 1–14, Grenoble, France, June 2012. ATALA/AFCP. URL http://www.aclweb.org/anthology/F12-2001.

Mishra K., Soni A., Sharma R., eta Sharma D. Exploring the Effects of Sentence Simplification on Hindi to English Machine Translation System. *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, 21–29, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/W14-5603.

Mitkov R. eta Štajner S. The Fewer, the Better? A Contrastive Study about Ways to Simplify. *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, 30–40, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/W14-5604.

Narayan S. eta Gardent C. Hybrid Simplification using Deep Semantics and Machine Translation. *the 52nd Annual Meeting of the Association for Computational Linguistics*, 435–445, 2014.

Narayan S. eta Gardent C. Unsupervised Sentence Simplification Using Deep Semantics. *arXiv preprint arXiv:1507.08452*, 2015.

Nunes B.P., Kawase R., Siehndel P., Casanova M.A., eta Dietze S. As Simple as It Gets - A Sentence Simplifier for Different Learning Levels and Contexts. *Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on*, 128–132, 2013.

Oh S.Y. Two Types of Input Modification and EFL Reading Comprehension: Simplification Versus Elaboration. *TESOL Quarterly*, 35(1):69–96, 2001.

Ondarra A. Murriztapen Gramatikaren sintaxia. EUSMG optimizatzen. Esaldi-mugak [The Syntax of Constraint Grammar. Optimising EUSMG. Clause Boundaries]. Master-tesia, University of the Basque Country (UPV/EHU), 2003.

Ong E., Damay J., Lojico G., Lu K., eta Tarantan D. Simplifying Text in Medical Literature. *J. Research in Science Computing and Eng*, 4(1): 37–47, 2007.

Paetzold G. Reliable Lexical Simplification for Non-Native Speakers. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 9–16, Denver, Colorado, June 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N15-2002.

Paetzold G. eta Specia L. LEXenstein: A Framework for Lexical Simplification. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, 85–90, Beijing, China, July 2015. Association for Computational Linguistics and The Asian Federation of Natural Language Processing. URL http://www.aclweb.org/anthology/P15-4015.

Paetzold G.H. eta Specia L. Text Simplification as Tree Transduction. In de Computaçao S.B., editor, *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 116–125, 2013.

Pang L.T. *Chinese Readability Analysis and its Applications on the Internet*. Doktoretza-tesia, The Chinese University of Hong Kong, 2006.

Pellow D. eta Eskenazi M. An Open Corpus of Everyday Documents for Simplification Tasks. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 84–93, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-1210.

Peng Y., Tudor C.O., Torii M., Wu C.H., eta Vijay-Shanker K. iSimp: A Sentence Simplification System for Biomedical Text. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1–6. IEEE, 2012.

Peng Y., Tudor C.O., Torii M., Wu C.H., eta Vijay-Shanker K. iSimp in BioC Standard Format: Enhancing the Interoperability of a Sentence Simplification System. *Database: The Journal of Biological Databases and Curation*, 2014. URL http://database.oxfordjournals.org/content/2014/bau038.abstract.

Petersen S.E. *Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education*. Doktoretza-tesia, Citeseer, 2007.

Petersen S.E. eta Ostendorf M. Text Simplification for Language Learners: A Corpus Analysis. *In Proceedings of Workshop on Speech and Language Technology for Education. SLaTE*, 69–72. Citeseer, 2007. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.9227&rep=rep1&type=pdf.

Petersen S.E. eta Ostendorf M. A Machine Learning Approach to Reading Level Assessment. *Computer Speech & Language*, 23(1):89–106, 2009.

Pitler E. eta Nenkova A. Revisiting Readability: A Unified Framework for Predicting Text Quality. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 186–195. Association for Computational Linguistics, 2008.

Platt J.C. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Schölkopf B., Burges C.J.C., eta Smola A.J., editors, *Advances in Kernel Methods-Support Vector Learning*. MIT Press, 1998. URL http://research.microsoft.com/~jplatt/smo.html.

Poornima C., Dhanalakshmi V., Anand K., eta Soman K. Rule based Sentence Simplification for English to Tamil Machine Translation System. *International Journal of Computer Applications*, 25(8):38–42, 2011.

Quinlan R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

Rello L. *DysWebxia. A Text Accessibility Model for People with Dyslexia*. Doktoretza-tesia, Universitat Pompeu Fabra, 2014.

Rello L., Baeza-Yates R., Bott S., eta Saggion H. Simplify or help?: Text simplification strategies for people with dyslexia. *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, 15:1–15:10, New York, NY, USA, 2013a. ACM. ISBN 978-1-4503-1844-0. URL http://doi.acm.org/10.1145/2461121.2461126.

Rello L., Baeza-Yates R., eta Saggion H. The Impact of Lexical Simplification by Verbal Paraphrases for People with and without Dyslexia. *Computational Linguistics and Intelligent Text Processing*, 501–512. Springer, 2013b.

Rello L., Bautista S., Baeza-Yates R., Gervás P., Hervás R., eta Saggion H. One Half or 50%? An Eye-Tracking Study of Number Representation Readability. *Proc. INTERACT*, 13 lib., 1–17, 2013c.

Rello L., Bayarri C., Baeza-Yates R., Gupta S., Kanvinde G., Saggion H., Bott S., Carlini R., eta Topac V. DysWebxia 2.0! More Accessible Text for People with Dyslexia. *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. ACM, 2013d.

Rennes E. eta Jönsson A. A Tool for Automatic Simplification of Swedish Texts. *Nordic Conference of Computational Linguistics NODALIDA 2015*, 317–320, 2015.

Rybing J., Smith C., eta Silvervarg A. Towards a Rule Based System for Automatic Simplification of texts. *The Third Swedish Language Technology Conference (SLTC 2010)*, 17–18, 2010.

Saggion H., Bott S., eta Rello L. Comparing Resources for Spanish Lexical Simplification. *SLSP 2013: 1st International Conference on Statistical Language and Speech Processing*, 1–12. Springer, 2013.

Saggion H., Bott S., eta Rello L. Simplifying Words in Context. Experiments with two Lexical Resources in Spanish. *Computer Speech & Language*, 35: 200 – 218, 2016. ISSN 0885-2308. URL http://www.sciencedirect.com/science/article/pii/S0885230815000078.

Saggion H., Gómez-Martínez E., Etayo E., Anula A., eta Bourg L. Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 47:341–342, 2011.

Saggion H., Marimon M., eta Ferrés D. Simplificación automática de textos para la accesibilidad de colectivos con discapacidad: experiencias para el español y el inglés. *IX Jornadas Científicas Internacionales de Investigación sobre Personas con Discapacidad*, 1–13, 2015a. ISBN 978-84-606-6434-5.

Saggion H., Štajner S., Bott S., Mille S., Rello L., eta Drndarevic B. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Trans. Access. Comput.*, 6(4):14:1–14:36, May 2015b. ISSN 1936-7228. URL http://doi.acm.org/10.1145/2738046.

Saini S., Sehgal U., eta Sahula V. Relative Clause Based Text Simplification for Improved English to Hindi Translation. *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, 1479–1484. IEEE, 2015.

Sato S., Matsuyoshi S., eta Kondoh Y. Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In Calzolari (Conference Chair) N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., eta Tapias D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 654–660, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. URL http://www.lrec-conf.org/proceedings/lrec2008/.

Scarton C., de Oliveira M., Candido Jr A., Gasperin C., eta Aluísio S.M. SIMPLIFICA: a Tool for Authoring Simplified Texts in Brazilian Portuguese Guided by Readability Assessments. *Proceedings of the NAACL HLT 2010 Demonstration Session*, 41–44. Association for Computational Linguistics, 2010.

Scarton C.E. eta Aluísio S.M. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamática*, 2(1):45–61, 2010.

Schwarm S.E. eta Ostendorf M. Reading Level Assessment using Support Vector Machines and Statistical Language Models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 523–530. Association for Computational Linguistics, 2005.

Seretan V. Acquisition of Syntactic Simplification Rules for French. In Calzolari (Conference Chair) N., Choukri K., Declerck T., Uğur Doğan M., Maegaard B., Mariani J., Odijk J., eta Piperidis S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 4019–426, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Shardlow M. Bayesian Lexical Simplification. Barne-txostena, Short Taster Research Project. The University of Manchester, 2012.

Shardlow M. A Comparison of Techniques to Automatically Identify Complex Words. *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, 103–109, Sofia, Bulgaria, August 2013a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P13-3015.

Shardlow M. The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, 69–77, Sofia, Bulgaria, August 2013b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2908.

Shardlow M. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing*, 58–70, 2014.

Sheremetyeva S. Automatic Text Simplification For Handling Intellectual Property (The Case of Multiple Patent Claims). *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, 41–52, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/W14-5605.

Si L. eta Callan J. A Statistical Model for Scientific Readability. *Proceedings of the tenth international conference on Information and knowledge management*, 574–576. ACM, 2001.

Siddharthan A. An Architecture for a Text Simplification System. *Proceedings of the Language Engineering Conference (LEC'02)*, 64–71, Washington, DC, USA, 2002. IEEE Computer Society. ISBN 0-7695-1885-0. URL http://dl.acm.org/citation.cfm?id=788016.788727.

Siddharthan A. *Syntactic Simplification and Text Cohesion*. Doktoretzatesia, University of Cambridge, 2003.

Siddharthan A. Syntactic Simplification and Text Cohesion. *Research on Language & Computation*, 4(1):77–109, 2006.

Siddharthan A. Complex Lexico-Syntactic Reformulation of Sentences using Typed Dependency Representations. *Proceedings of the 6th International Natural Language Generation Conference*, 125–133. Association for Computational Linguistics, 2010.

Siddharthan A. Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. *Proceedings of the 13th European Workshop on Natural Language Generation*, 2–11. Association for Computational Linguistics, 2011.

Siddharthan A. A Survey of Research on Text Simplification. *The International Journal of Applied Linguistics*, 259–98, 2014.

Siddharthan A., Nenkova A., eta McKeown K. Syntactic Simplification for Improving Content Selection in Multi-Document Summarization. *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004.

Silveira Botelho S. eta Branco A. Enhancing Multi-document Summaries with Sentence Simplification. *ICAI 2012: International Conference on Artificial Intelligence*, 2012.

Simensen A.M. Adapted Readers: How are they Adapted. *Reading in a Foreign Language*, 4(1):41–57, 1987.

Sinha M., Sharma S., Dasgupta T., eta Basu A. New Readability Measures for Bangla and Hindi Texts. *Proceedings of COLING 2012: Posters*, 1141–1150, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL http://www.aclweb.org/anthology/C12-2111.

Sinha R. UNT-SimpRank: Systems for Lexical Simplification Ranking. *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 493–496. Association for Computational Linguistics, 2012.

Sjöholm J. Probability as Readability: A New Machine Learning Approach to Readability Assessment for Written Swedish. Master-tesia, Linköping, 2012.

Soraluze A., Arregi O., Arregi X., eta Díaz de Ilarraza A. Coreference Resolution for Morphologically Rich Languages. Adaptation of the Stanford System to Basque. 55:23–30, 2015.

Soraluze A., Arregi O., Arregi X., Klara Ceberio K., eta Díaz de Ilarraza A. Mention Detection: First Steps in the Development of a Basque Coreference Resolution System. *Proceedings of KONVENS 2012 (Main track: oral presentations)*, 128–163, 2012.

Specia L. Translating from Complex to Simplified Sentences. *Computational Processing of the Portuguese Language*, 30–39, 2010.

Specia L., Aluísio S.M., eta Pardo T.A. Manual de Simplificação Sintática para o Português. Barne-txostena NILC-TR-08-06, São Carlos-SP., 2008.

Specia L., Jauhar S.K., eta Mihalcea R. Semeval-2012 Task 1: English Lexical Simplification. *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, 347–355, 2012.

Srivastava J. eta Sanyal S. Segmenting Long Sentence Pairs to Improve Word Alignment in English-Hindi Parallel Corpora. *Advances in Natural Language Processing*, 97–107. Springer, 2012.

Štajner S. Translating Sentences from 'Original' to 'Simplified' Spanish. *Procesamiento del Lenguaje Natural*, 53:61–68, 2014.

Štajner S. *New Data-Driven Approaches to Text Simplification*. Doktoretza-tesia, University of Wolverhampton, 2015.

Štajner S., Béchara H., eta Saggion H. A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 823–828, 2015a.

Štajner S., Calixto I., eta Saggion H. Automatic Text Simplification for Spanish: Comparative Evaluation of Various Simplification Strategies. *Proceedings of Recent Advances in Natural Language Processing*, 618–626, 2015b.

Štajner S., Drndarevic B., eta Saggion H. Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification. *Computacion y Systemas*, 17(2):251–262, 2013.

Štajner S., Evans R., eta Dornescu I. Assessing Conformance of Manually Simplified Corpora with User Requirements: the Case of Autistic Readers. *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, 53–63, Dublin, Ireland, August 2014a. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/W14-5606.

Štajner S., Mitkov R., eta Pastor G.C. Simple or Not Simple? A Readability Question. *Language Production, Cognition, and the Lexicon*, 379–398, 2015c.

Štajner S., Mitkov R., eta Saggion H. One Step Closer to Automatic Evaluation of Text Simplification Systems. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 1–10, Gothenburg, Sweden, April 2014b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-1201.

Štajner S. eta Saggion H. Readability Indices for Automatic Evaluation of Text Simplification Systems: A Feasibility Study for Spanish. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 374–382, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. URL http://www.aclweb.org/anthology/I13-1043.

Štajner S. eta Saggion H. Translating from Original to Simplified Sentences using Moses: When does it Actually Work? *Proceedings of Recent Advances in Natural Language Processing*, 611–617, 2015.

Stenetorp P., Pyysalo S., Topic G., Ohta T., Ananiadou S., eta Tsujii J. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. *Proceedings of the Demonstrations Session at EACL 2012*, 102–107, 2012.

Suter J. Rule-based Text Simplification for German. Master-tesia, Universität Zürich, 2015.

Temnikova I. *Text Complexity and Text Simplification in the Crisis Management Domain*. Doktoretza-tesia, University of Wolverhampton, 2012.

Temnikova I. eta Maneva G. The C-Score – Proposing a Reading Comprehension Metrics as a Common Evaluation Measure for Text Simplification. *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, 20–29, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2903.

Temnikova I., Orasan C., eta Mitkov R. CLCM - A Linguistic Resource for Effective Simplification of Instructions in the Crisis Management Domain and its Evaluations. In Calzolari (Conference Chair) N., Choukri K., Declerck T., Uğur Doğan M., Maegaard B., Mariani J., Odijk J., eta Piperidis S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 3007–3014, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Tesnière L. *Eléments de syntaxe structurale*. Librairie C. Klincksieck, Paris, 1959.

Thomas S.R. eta Anderson S. WordNet-Based Lexical Simplification of a Document. *Empirical Methods in Natural Language Processing*, 80–88, 2012.

Tur G., Hakkani-Tur D., Heck L., eta Parthasarathy S. Sentence Simplification for Spoken Language Understanding. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5628–5631. IEEE, 2011.

Tyagi S., Chopra D., Mathur I., eta Joshi N. Classifier based Text Simplification for improved Machine Translation. *Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances*, 46–50. IEEE, 2015.

Urizar R. *Euskal lokuzioen tratamendu konputazionala [Computational Treatment of Basque Locutions]*. Doktoretza-tesia, University of the Basque Country (UPV/EHU), 2012.

Urkia M. *Euskal morfologiaren tratamendu informatikorantz [Towards the Computational Treatment of Basque Morphology]*. Doktoretza-tesia, University of the Basque Country (UPV/EHU), 1997.

Vajjala S. eta Meurers D. Assessing the Relative Reading Level of Sentence Pairs for Text Simplification. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 288–297, Gothenburg, Sweden, April 2014a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E14-1031.

Vajjala S. eta Meurers D. Readability Assessment for Text Simplification: From Analysing Documents to Identifying Sentential Simplifications. *International Journal of Applied Linguistics*, 165(2):194–222, 2014b.

Vanderwende L., Suzuki H., Brockett C., eta Nenkova A. Beyond SumBasic: Task-focused Summarization with Sentence Simplification and Lexical Expansion. *Information Processing & Management*, 43(6):1606–1618, 2007.

Vertan C. eta von Hahn W. Making Historical Texts Accessible to Everybody. *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, 64–68, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/W14-5607.

Vettori C. eta Mich O. Supporting Deaf Children's Reading Skills: The Many Challenges of Text Simplification. *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, ASSETS '11, 283–284, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0920-2. URL http://doi.acm.org/10.1145/2049536.2049608.

Vickrey D. eta Koller D. Sentence Simplification for Semantic Role Labeling. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2008: HLT)*, 344–352, 2008.

vor der Brück T., Hartrumpf S., eta Helbig H. A Readability Checker with Supervised Learning Using Deep Indicators. *Informatica*, 32(4):429–435, 2008.

Vu T.T., Tran G.B., eta Pham S.B. Learning to Simplify Children Stories with Limited Data. *Intelligent Information and Database Systems*, 31–41. Springer, 2014.

Wei C.H., Leaman R., eta Lu Z. SimConcept: a Hybrid Approach for Simplifying Composite Named Entities in Biomedicine. *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 138–146. ACM, 2014.

Wilkens R., Dalla Vecchia A., Boito M.Z., Padró M., eta Villavicencio A. Size Does Not Matter. Frequency Does. A Study of Features for Measuring Lexical Complexity. In Bazzan A.L. eta Pichara K., editors, *Advances in Artificial Intelligence – IBERAMIA 2014*, 8864 lib. of *Lecture Notes in Computer Science*, 129–140. Springer International Publishing, 2014. ISBN 978-3-319-12026-3. URL http://dx.doi.org/10.1007/978-3-319-12027-0_11.

Woodsend K. eta Lapata M. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 409–420, Stroudsburg, PA, USA, 2011a. Association for Computational Linguistics.

Woodsend K. eta Lapata M. WikiSimple: Automatic Simplification of Wikipedia Articles. *Proceedings of the TwentyFifth AAAI Conference on Artificial Intelligence*, 927–932, 2011b. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/download/3505/3968.

Wubben S., van den Bosch A., eta Krahmer E. Sentence Simplification by Monolingual Machine Translation. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 1015–1024. Association for Computational Linguistics, 2012.

Xu W., Callison-Burch C., eta Napoles C. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297, 2015.

Yatskar M., Pang B., Danescu-Niculescu-Mizil C., eta Lee L. For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplifications from Wikipedia. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, 365–368, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL http://dl.acm.org/citation.cfm?id=1857999.1858055.

Young D.N. Linguistic Simplification of SL Reading Material: Effective Instructional Practice? *The Modern Language Journal*, 83(3):350–366, 1999.

Zamanian M. eta Heydari P. Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2(1):43–53, 2012.

Zhu Z., Bernhard D., eta Gurevych I. A Monolingual Tree-based Translation Model for Sentence Simplification. *Proceedings of The 23rd International Conference on Computational Linguistics*, 1353–1361. Association for Computational Linguistics, 2010.

A

Perpaus adberbialen egiturak

Eranskin honetan perpaus adberbialen egituren zerrenda aurkezten dugu. Egiturak perpaus-motaren arabera zerrendatu ditugu.

Denbora jokatu
-enean
-ela
-elarik
Noiz eta...bait-/-en
-enetan
-en bakoitzean
-en guztietan
-en aldikal
-en aldiro
Zenbat aldiz -en... hainbat aldiz
-eneko
-en orduko
-en bezain laster
-en bezain sarri
-en bezain agudo
-en bezain fite
-en ber
-enaz batera
-en baino lehen
-en aurrean
-en aitzinean
-en ondoan
-en ostean
-enetik
-enez gero
-enik...-ra
-en arte
-eno
-eino
-en bitartean
-en artean
-en arteko
-enerako
-en heinean
-en momentuan

Denbora ezjokatu
-tzean
-tzerakoan
-tzekoan
-tzearekin
-tzeari
-tzerat
-tu(k)eran
-tu aldiro
-tu bakoitzean
-tu guztian
-tu ahala
-tu arau
-tzerako
-tu orduko
-tu bezain laster
-tu bezain pronto
-tu eta berehala
-tu eta laster
-tuaz
-tzearekin bat(era)
-tu berri(t)an
-tu ahala
-tu arau
-tu baino lehen
-tu aurretik
-tu aitzinean
-tu gabe
-tu eta
-tu eta gero
-tuta
-tutakoan
-tu ondoren
-tu ondoan
-tu ostean
-tu(a)z
-tuz gero
-turik
-tu arte
-tu artean

-tu bitartean
-tzeraino
-tu aitzin
-tu osteko
-tu bezperan
-tu ondotik
-tzear

Kausa jokatu

-elako((t)z)
-elakoan
bait-
... eta
zeren eta ...(bait-/-(e)n)
zeren
zergatik
Ezen (...) (bait-)
-enez gero(ztik)
-enez
-en legez
nola/zelan... (-en/bait)
-ela kausa
-ela bide
-ela medio

Kausa ezjokatu

-tzeagatik
-tzearren
-turik
-tutakoan
-tuz
-tuta
-tuz gero
-tzearekin

Kontzesio jokatu

Nahiz (eta) -(en/-ela/ba-)
-en arren
ba- (...) ere

Kontzesio ezjokatu

nahiz eta ... -tu/0
-tu arren
-tugatik
-tuta (gabe/ezta) ere
-turik (gabe/ezta) ere
-tuz gero ere
-tzearren
-tuz gero
-ik ere

Modu jokatu

-ela
-elarik
-en moduan/ra
-en arabera(n)
-en eran/ra
-en antzera
-en moldean/ra
-en gisan/ra
-en bezala
-en legez
-enez

Modu ezjokatu

-tuz
-tuta
-turik
-tu gabe
-tu barik
-tu ezta
-tu ordez
-tu ordean
-tu aginean
-tu aginik
-tu ahala
-tu arau
-tu beharrean
-tu nahirik
-tu nahian
-tu gurarik
-tu ezinik
-tu ezinda
-tu ezinean
-tu beharrez
-tzeko zorian
-tu hurran
-tzeko moduan
-tzeko gisan
-tzeko eran
-tzeko maneran
-tzeko moldean
-tzekotan
-tu bezala
-tzeke
-era
-tzeaz
-tzeaz gain
-tik

Helburu jokatu

subjuntiboa

Helburu ezjokatu

-tzeko(tz)
-tzekotzat
-tzearren
-tzeagatik
-tze alde(ra)
-tzekotan
-tzeko asmotan
-tzeko intentziotan
-tzeko intentzioarekin
-tzera

Ondorio jokatu

(...) (non) ... bait-
(...) (ezen) ... -en

Baldintza jokatu

ba-
non ez... -(e)n

Baldintza ezjokatu

-tuz gero((z)tik)
-tu ezkero(an/k/tino/((z)tik))
-tu ezik
-tu ezean
-tzekotan
-tzekoz
-tzez gero
-tuenean
-tzera(t)

Egitura konplexuak sinplifikatzeko erregelak

Eranskin honetan 3. kapituluan aurkeztutako sinplifikazio-proposamenen erregelak aurkeztuko ditugu. Taulak fenomenoen arabera egin ditugu (koordinazioa B.1 taulan, perpaus erlatiboak B.2 taulan, perpaus osagarriak B.3 taulan, aposizio-sintagmak B.4 taulan, egitura parentetikoak B.5 taulan, eta perpaus adberbialak B.6 taulan). Taulak honela antolatu ditugu: mota zutabean aztergai dugun egituraren mota (denbora, konpletiboa,...) zein den; egitura zutabean, lantzen dugun egitura; ezabatzeko (ezaba.) zutabean, ezabatu behar diren erlazio-markak (Erlazio_Marken_Zerrenda); txertatzeko (txert.) zutabean, txertatze-elementu lehenetsia; Txert2, Txert3eta Txert4 zutabeetan, egonez gero, txertatze-elementu alternatiboak, txertatu non? zutabean, txertatzeko elementuak non txertatu behar diren (Txertatze_Elementuen_Zerrenda); hurrenkera (hurrenk.) zutabean, esaldi berrien hurrenkera (Hurrenkeren_Zerrenda); eta oharrak zutabean, oharrak jarri ditugu.

Mota	**Egitura**	**Ezabatzeko**	**Txertatzeko**	**Txertatu non?**	**Hurrenkera**	**Oharrak**
Emendioa	eta	eta	ø		$koord_1$-$koord_2$	
Aurkaritza	baina	baina	Baina	$koord_2$	$koord_1$-$koord_2$	
Hautakaritza	edo	edo	Edo	$koord_2$	$koord_1$-$koord_2$	
Alborakuntza	;	;	ø		$koord_1$-$koord_2$	
Alborakuntza	,	,	ø		$koord_1$-$koord_2$	

B.1 taula – Euskarazko egiturak sinplifikatzeko erregelak (koordinazioa)

Mota	**Egitura**	**Ezabatzeko**	**Txertatzeko**	**Txertatu non?**	**Hurrenkera**	**Oharrak**
Arruntak	-en	-en	aurrekaria + determinatzailea	nagusiaren hasieran	mendeko-a_{jat}-nagusi-a_{jat}	kasu-markak egokitu; determinatzailerik badago, ez txertatu; aurrekaria entitatea bada, determinatzailea ez txertatu

(Jarraipena hurrengo orrialdean)

Mota	Egitura	Ezabatzeko	Txertatzeko	Txertatu non?	Hurrenkera	Oharrak
Zein erlatiboak	zein	zein	aurrekaria + determinatzaile	mendekoaren hasieran	nagusia$_{jat}$-mendekoa$_{jat}$	kasu-markak egokitu

B.2 taula – Euskarazko egiturak sinplifikatzeko erregelak (perpaus erlatiboak)

Mota	Egitura	Ezaba.	Txert.	Txert2	Txertatu non?	Hurrenk.	Oharrak
Konpletiboak	-ela	-ela	honako hau	ø	nagusian, aditzaren aurretik	nagusia$_{jat}$-mendekoa$_{jat}$	izenordainen aldaketa; aditzaren pertsonen aldaketa; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean
Konpletiboak	-enik	-enik	honako hau	ø	nagusian, aditzaren aurretik	nagusia$_{jat}$-mendekoa$_{jat}$	izenordainen aldaketa; aditzaren pertsonen aldaketa; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Konpletiboak	-tzen	-tzen	honako hau	ø	nagusian, aditzaren aurretik	nagusia$_{jat}$-mendeko-a$_{jat}$	predikatibo funtzioa badute, ez sinplifikatu; perpaus nagusia bi punturekin amaitu, mendeko perpausa letra xehez hasi
Konpletiboak	-tzera	-tzera	honako hau	ø	nagusian, aditzaren aurretik	nagusia$_{jat}$-mendeko-a$_{jat}$	predikatibo funtzioa badute, ez sinplifikatu; perpaus nagusia bi punturekin amaitu, mendeko perpausa letra xehez hasi
Konpletiboak	-tzeko	-tzeko	honako hau	ø	nagusian, aditzaren aurretik	nagusia$_{jat}$-mendeko-a$_{jat}$	predikatibo funtzioa badute, ez sinplifikatu; perpaus nagusia bi punturekin amaitu, mendeko perpausa letra xehez hasi
Konpletiboak	-tzea + atzizkia	-tzea + atzizkia	honako hau	ø	nagusian, aditzaren aurretik	nagusia$_{jat}$-mendeko-a$_{jat}$	predikatibo funtzioa badute, ez sinplifikatu; perpaus nagusia bi punturekin amaitu, mendeko perpausa letra xehez hasi
Konpletiboak	-tzeari	-tzeari	honako honi	ø	nagusian, aditzaren aurretik	nagusia$_{jat}$-mendeko-a$_{jat}$	perpaus nagusia bi punturekin amaitu, mendeko perpausa letra xehez hasi

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Konpletiboak	-tzerik	-tzerik	honako hau	ø	nagusian, aditzaren aurretik	nagusia$_{jat}$-mendeko-a$_{jat}$	predikatibo funtzioa badute, ez sinplifikatu; perpaus nagusia bi punturekin amaitu, mendeko perpausa letra xehez hasi
Konpletiboak	-tu izana	-tu izana	honako hau	ø	nagusian, aditzaren aurretik	nagusia$_{jat}$-mendeko-a$_{jat}$	predikatibo funtzioa badute, ez sinplifikatu; perpaus nagusia bi punturekin amaitu, mendeko perpausa letra xehez hasi
Konpletiboak	-tu izanari	-tu izanari	honako honi	ø	nagusian, aditzaren aurretik	nagusia$_{jat}$-mendeko-a$_{jat}$	perpaus nagusia bi punturekin amaitu, mendeko perpausa letra xehez hasi
Zehar-galderak	-en	-en	honako hau	ø	nagusian, aditzaren aurretik	nagusia$_{jat}$-mendeko-a$_{jat}$	ea partikulua badago, hori ere ezabatu; izenordainen aldaketa; aditzaren pertsonen aldaketa; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Zehar-galderak	galde-tzaileak ... -en	-en	honako hau	ø	nagusian, aditzaren aurretik	nagusia$_{jat}$-mendeko-a$_{jat}$	galdetzaileak mendeko perpausaren hasieran jarri; ea partikulua badago, hori ere ezabatu; izenordainen aldaketa; aditzaren pertsonen aldaketa; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean
Zehar-galderak	ez-jokatuak	ø	ø	ø	ø	ø	EZ SINPLIFIKATU
Bestelakoak	-enez + aditz diskur-tsiboa	-enez	honako hau	ø	nagusian, aditzaren aurretik	mendeko-a$_{jat}$-nagu-sia$_{jat}$	izenordainen aldaketa; aditzaren pertsonen aldaketa: puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Postposizio-sintagmak	-en arabera	-en arabera	honako hau	dio/te	nagusian, aditzaren aurretik; txert2 perpaus amaieran	mendeko-a$_{jat}$-nagusia$_{jat}$	genitiboan dagoen hitza biziduna izan behar da; ergatiboa txertatu genitiboa kendu zaion hitzari; sing/pl kontuan izan aditza aukeratzeko; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean
Postposizio-sintagmak	-en araberan	-en araberan	honako hau	dio/te	nagusian, aditzaren aurretik; txert2 perpaus amaieran	mendeko-a$_{jat}$-nagusia$_{jat}$	genitiboan dagoen hitza biziduna izan behar da; ergatiboa txertatu genitiboa kendu zaion hitzari; sing/pl kontuan izan aditza aukeratzeko; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Postposizio-sintagmak	-en arau	-en arau	honako hau	dio/te	nagusian, aditzaren aurretik; txert2 perpaus amaieran	mendeko-a_{jat}-nagu-sia$_{jat}$	genitiboan dagoen hitza biziduna izan behar da; ergatiboa txertatu genitiboa kendu zaion hitzari; sing/pl kontuan izan aditza aukeratzeko; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean
Postposizio-sintagmak	-en arauaz	-en arauaz	honako hau	dio/te	nagusian, aditzaren aurretik; txert2 perpaus amaieran	mendeko-a_{jat}-nagu-sia$_{jat}$	genitiboan dagoen hitza biziduna izan behar da; ergatiboa txertatu genitiboa kendu zaion hitzari; sing/pl kontuan izan aditza aukeratzeko; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Postposizio-sintagmak	-en arauka	-en arauka	honako hau	dio/te	nagusian, aditzaren aurretik; txert2 perpaus amaieran	mendeko-a$_{jat}$-nagusia$_{jat}$	genitiboan dagoen hitza biziduna izan behar da; ergatiboa txertatu genitiboa kendu zaion hitzari; sing/pl kontuan izan aditza aukeratzeko; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean
Postposizio-sintagmak	-en ereduz	-en ereduz	honako hau	dio/te	nagusian, aditzaren aurretik; txert2 perpaus amaieran	mendeko-a$_{jat}$-nagusia$_{jat}$	genitiboan dagoen hitza biziduna izan behar da; ergatiboa txertatu genitiboa kendu zaion hitzari; sing/pl kontuan izan aditza aukeratzeko; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Postposizio-sintagmak	-en eredura	-en eredura	honako hau	dio/te	nagusian, aditzaren aurretik; txert2 perpaus amaieran	mendeko-a_{jat}-nagusia$_{jat}$	genitiboan dagoen hitza biziduna izan behar da; ergatiboa txertatu genitiboa kendu zaion hitzari; sing/pl kontuan izan aditza aukeratzeko; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean
Postposizio-sintagmak	-en hitzetan	-en hitzetan	honako hau	esan du/-te	nagusian, aditzaren aurretik; txert2 perpaus amaieran	mendeko-a_{jat}-nagusia$_{jat}$	genitiboan dagoen hitza biziduna izan behar da; ergatiboa txertatu genitiboa kendu zaion hitzari; sing/pl kontuan izan aditza aukeratzeko; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean

(Jarraipena hurrengo orrialdean)

Mota	Egitura	Ezaba.	Txert.	Txert2	Txertatu non?	Hurrenk.	Oharrak
Postposizio-sintagmak	-en adieraz-penetan	-en adie-razpenetan	honako hau	adierazi du/te	nagusian, aditzaren aurretik; txert2 perpaus amaieran	mendeko-a_{jat}-nagu-sia$_{jat}$	genitiboan dagoen hitza biziduna izan behar da; ergatiboa txertatu genitiboa kendu zaion hitzari; sing/pl kontuan izan aditza aukeratzeko; puntuazio-markak: nagusia bi punturekin amaitu eta mendekoa komatxoen artean

B.3 taula – Euskarazko egiturak sinplifikatzeko erregelak (perpaus osagarriak, bestelakoak eta postposizio-sintagmak)

Mota	Egitura	Ezabatzeko	Txert.	Txertatu non?	Hurrenkera	Oharrak
IS barnean	Entitatea + hon-darkia	Beharrezkoak ez diren kasu markak	da/dira	mendeko perpau-sean	nagusia$_{jat}$-mendekoa$_{jat}$	esaldi berrien barne-hurrenke-ra: entitateHon-darkiAditza; sing/pl kontuan izan aditza aukeratzeko

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezabatzeko**	**Txert.**	**Txertatu non?**	**Hurrenkera**	**Oharrak**
IS barnean	Hondarkia + entitatea	Beharrezkoak ez diren kasu markak	da/dira	mendeko perpausean	nagusia$_{jat}$-mendekoa$_{jat}$	esaldi berrien barne-hurrenkera: entitateHondarkiAditza; sing/pl kontuan izan aditza aukeratzeko
IS aposatuak		Beharrezkoak ez diren kasu markak	da/dira	mendeko perpausean	nagusia$_{jat}$-mendekoa$_{jat}$	esaldi berrien barne-hurrenkera: entitate-hondarki-aditza; sing/pl kontuan izan aditza aukeratzeko

B.4 taula – Euskarazko egiturak sinplifikatzeko erregelak (aposizio-sintagmak)

Mota	**Egitura**	**Ezabatzeko**	**Txert.**	**Txert2**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Toki informazioa		inesiboa nagusian	dago/daude	inesiboa	mendeko perpausean	nagusia$_{jat}$-mendeko-a$_{jat}$	esaldi berrien barne hurrenkera: tokiaTokiparentetikoaAditza; sing/pl kontuan izan aditza aukeratzeko

(Jarraipena hurrengo orrialdean)

Mota	Egitura	Ezabatzeko	Txert.	Txert2	Txertatu non?	Hurrenk.	Oharrak
Jaiotza datuak		Beharrezkoak ez diren kasu markak	jaio zen	inesiboa	mendeko perpausean	nagusia$_{jat}$-mendeko-a$_{jat}$	esaldi berrien barne hurrenkera: entitateaJaiotzeegu-na,JaiotzetokiaAditza; inesiboa tokian eta datan gehitu
Heriotz datuak		Beharrezkoak ez diren kasu markak	hil zen	inesiboa	mendeko perpausean	nagusia$_{jat}$-mendeko-a$_{jat}$	esaldi berrien barne hurrenkera: entitateaHeriotzegu-na,HeriotztokiaAditza; inesiboa tokian eta datan gehitu

B.5 taula – Euskarazko egiturak sinplifikatzeko erregelak (egitura parentetikoak)

Mota	Egitura	Ezaba.	Txert.	Txert2	Txert3	Txert4	Txertatu non?	Hurrenk.	Oharrak
Denbora	-enean	-enean	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-ela	-ela	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-elarik	-elarik	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Denbora	noiz eta... bait-	noiz eta... bait-	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	noiz eta... -en	noiz eta... -en	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-tzean	-tzean	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	- tzera-koan	- tzera-koan	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-tzekoan	-tzekoan	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	- tzeare-kin	- tzeare-kin	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-tzeari	-tzeari	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-tzerat	-tzerat	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Denbora	-tueran	-tueran	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-tukeran	-tukeran	Orduan	Une hartan	Aldi berean	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-enetan	-enetan	Une horietan guztietan	Aldiro	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-en bakoitzean	-en bakoitzean	Une horietan guztietan	Aldiro	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-en guztietan	-en guztietan	Une horietan guztietan	Aldiro	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-en aldikal	-en aldikal	Une horietan guztietan	Aldiro	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-en aldiro	-en aldiro	Une horietan guztietan	Aldiro	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	zenbat aldiz -en... hainbat aldiz	zenbat aldiz -en... hainbat aldiz	Une horietan guztietan	Aldiro	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Denbora	-tu aldiro	-tu aldiro	Une horietan guztietan	Aldiro	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-tu bakoitzean	-tu bakoitzean	Une horietan guztietan	Aldiro	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-tu guztian	-tu guztian	Une horietan guztietan	Aldiro	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-tu ahala	-tu ahala	Une horietan guztietan	Aldiro	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-tu arau	-tu arau	Une horietan guztietan	Aldiro	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-eneko	-eneko	Orduko	Segidan	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-en orduko	-en orduko	Orduko	Segidan	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Denbora	-tzerako	-tzerako	Orduko	Segidan	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Denbora	-tu orduko	-tu orduko	Orduko	Segidan	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-en bezain laster	-en bezain laster	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-en bezain sarri	-en bezain sarri	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-en bezain agudo	-en bezain agudo	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-en bezain fite	-en bezain fite	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-en ber	-en ber	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-enaz batera	-enaz batera	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-tu bezain laster	-tu bezain laster	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Denbora	-tu bezain pronto	-tu bezain pronto	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-tu eta berehala	-tu eta berehala	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-tu eta laster	-tu eta laster	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-tuaz	-tuaz	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	- tzearekin bat	- tzearekin bat	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	- tzearekin batera	- tzearekin batera	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-tu berrian	-tu berrian	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-tu berritan	-tu berritan	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Denbora	-tu ahala	-tu ahala	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-tu arau	-tu arau	Une horretan bertan	Orduko	Segidan	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-en baino lehen	-en baino lehen	Gero	Ondoren	Ostean	ø	nagusiaren hasieran	nagusi-a$_{jat}$-mendekoa$_{jat}$	
Denbora	-en aurrean	-en aurrean	Gero	Ondoren	Ostean	ø	nagusiaren hasieran	nagusi-a$_{jat}$-mendekoa$_{jat}$	
Denbora	-en aitzinean	-en aitzinean	Gero	Ondoren	Ostean	ø	nagusiaren hasieran	nagusi-a$_{jat}$-mendekoa$_{jat}$	
Denbora	-tu baino lehen	-tu baino lehen	Gero	Ondoren	Ostean	ø	nagusiaren hasieran	nagusi-a$_{jat}$-mendekoa$_{jat}$	
Denbora	-tu aurretik	-tu aurretik	Gero	Ondoren	Ostean	ø	nagusiaren hasieran	nagusi-a$_{jat}$-mendekoa$_{jat}$	
Denbora	-tu aintzinean	-tu aintzinean	Gero	Ondoren	Ostean	ø	nagusiaren hasieran	nagusi-a$_{jat}$-mendekoa$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Denbora	-tu gabe	-tu gabe	Gero	Ondoren	Ostean	ø	nagusiaren hasieran	nagusi-a_{jat}-mendekoa$_{jat}$	
Denbora	-tu orduko	-tu orduko	Gero	Ondoren	Ostean	ø	nagusiaren hasieran	nagusi-a_{jat}-mendekoa$_{jat}$	ANB
Denbora	-tzerako	-tzerako	Gero	Ondoren	Ostean	ø	nagusiaren hasieran	nagusi-a_{jat}-mendekoa$_{jat}$	ANB
Denbora	-en ondoan	-en ondoan	Ondoren	Ostean	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-en ondoren	-en ondoren	Ondoren	Ostean	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-en ostean	-en ostean	Ondoren	Ostean	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-tu eta	-tu eta	Ondoren	Ostean	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Denbora	-tu eta gero	-tu eta gero	Ondoren	Ostean	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Denbora	-tuta	-tuta	Ondoren	Ostean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-tu ondoan	-tu ondoan	Ondoren	Ostean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-tu ostean	-tu ostean	Ondoren	Ostean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-tuz	-tuz	Ondoren	Ostean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-tuaz	-tuaz	Ondoren	Ostean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-tuz gero	-tuz gero	Ondoren	Ostean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-turik	-turik	Ondoren	Ostean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-enetik	-enetik	Ordutik	Une hartatik	Harrezkero	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Denbora	-enez gero	-enez gero	Ordutik	Une hartatik	Harrezkero	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-enik -ra	-enik -ra	Ordutik	Une hartatik	Harrezkero	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-tuz gero	-tuz gero	Ordutik	Une hartatik	Harrezkero	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	ANB
Denbora	-en arte	-en arte	Ordura arte	Orduraino	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-tu arte	-tu arte	Ordura arte	Orduraino	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-tu bitartean	-tu bitartean	Ordura arte	Orduraino	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-tzeraino	-tzeraino	Ordura arte	Orduraino	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	
Denbora	-eno	-eno	Bitartean	Artean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-nagusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Denbora	-eino	-eino	Bitartean	Artean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	
Denbora	-en bitar-tean	-en bitar-tean	Bitartean	Artean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	
Denbora	-en ar-tean	-en ar-tean	Bitartean	Artean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	
Denbora	-en arte-ko	-en arte-ko	Bitartean	Artean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	
Denbora	-tu bitar-te	-tu bitar-te	Bitartean	Artean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	ANB
Denbora	-tu bitar-tean	-tu bitar-tean	Bitartean	Artean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	ANB
Denbora	-tu ar-tean	-tu ar-tean	Bitartean	Artean	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	ANB
Kausa	-elako	-elako	Horregatik	Hori del-eta	ø	ø	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Kausa	-elakoz	-elakoz	Horregatik	Hori dela-eta	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kausa	-elakotz	-elakotz	Horregatik	Hori dela-eta	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kausa	-elakoan	-elakoan	Horregatik	Hori dela-eta	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kausa	bait-	bait-	Horregatik	Hori dela-eta	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kausa	zeren eta ... bait-	zeren eta ... bait-	Horregatik	Hori dela-eta	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kausa	zeren eta ... -en	zeren eta ... -en	Horregatik	Hori dela-eta	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kausa	- tzeaga-tik	- tzeaga-tik	Horregatik	Hori dela-eta	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kausa	-tzearren	-tzearren	Horregatik	Hori dela-eta	ø	ø	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Kausa	bait-	bait-	Izan ere	ø	ø	ø	nagusiaren hasieran	nagusi-a_{jat}-mendekoa$_{jat}$	Azalpenezkoak
Kausa	... eta	... eta	Izan ere	ø	ø	ø	nagusiaren hasieran	nagusi-a_{jat}-mendekoa$_{jat}$	Azalpenezkoak
Kausa	zeren eta ... bait-	zeren eta ... bait-	Izan ere	ø	ø	ø	nagusiaren hasieran	nagusi-a_{jat}-mendekoa$_{jat}$	Azalpenezkoak
Kausa	zeren eta ... -en	zeren eta ... -en	Izan ere	ø	ø	ø	nagusiaren hasieran	nagusi-a_{jat}-mendekoa$_{jat}$	Azalpenezkoak
Kontzesioa	nahiz eta... - en	nahiz eta - en	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Kontzesioa	nahiz eta... - ela	nahiz eta - ela	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Kontzesioa	nahiz eta... ba-	nahiz eta ba-	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Kontzesioa	nahiz... -en	nahiz - en	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kontzesioa	nahiz... -ela	nahiz - ela	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kontzesioa	nahiz... ba-	nahiz ba-	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kontzesioa	-en arren	-en arren	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kontzesioa	ba- ere	ba- ere	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kontzesioa	nahiz eta -tu	nahiz eta -tu	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Kontzesioa	-tu arren	-tu arren	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	
Kontzesioa	-tuagatik	-tuagatik	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	
Kontzesioa	-tuta ere	-tuta ere	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	
Kontzesioa	-tuta gabe ere	-tuta gabe ere	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	
Kontzesioa	-tuta ezta ere	-tuta ezta ere	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	
Kontzesioa	-tzearren	-tzearren	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a$_{jat}$-na-gusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Kontzesioa	-tuz gero	-tuz gero	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Kontzesioa	-ik ere	-ik ere	Hala ere	Nolanahi ere	Edonola ere	Hala eta guztiz ere	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Modua	-ela	-ela	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Modua	-elarik	-elarik	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Modua	-en moduan	-en moduan	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Modua	-en modura	-en modura	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Modua	-en antzera	-en antzera	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	
Modua	-en bezala	-en bezala	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Modua	-en legez	-en legez	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Modua	-enez	-enez	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	aditz diskurtsiborik gabe
Modua	-tuz	-tuz	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Modua	-tuta	-tuta	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Modua	-turik	-turik	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	
Modua	-tu gabe	-tu gabe	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	polaritate aldaketa mendeko perpausean; BEREZI
Modua	-tu barik	-tu barik	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	polaritate aldaketa mendeko perpausean; BEREZI
Modua	-tu ezta	-tu ezta	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	polaritate aldaketa mendeko perpausean; BEREZI

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Modua	-tu ordez	-tu ordez	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	polaritate aldaketa mendeko perpausean; BEREZI
Modua	-tu ordean	-tu ordean	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	polaritate aldaketa mendeko perpausean; BEREZI
Modua	-tzeke	-tzeke	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	polaritate aldaketa mendeko perpausean; BEREZI
Modua	-tu beharrean	-tu beharrean	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	polaritate aldaketa mendeko perpausean; BEREZI
Modua	-tu aginean	-tu aginean	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	mendeko perpausean ia txertatu; BEREZI
Modua	-tu aginik	-tu aginik	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-nagusia$_{jat}$	mendeko perpausean ia txertatu; BEREZI

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Modua	-tzeko zorian	-tzeko zorian	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	mendeko perpausean ia txertatu; BEREZI
Modua	-tu hurran	-tu hurran	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	mendeko perpausean ia txertatu; ANB; BEREZI
Modua	-tu beharrean	-tu beharrean	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	mendeko perpausean behar izan txertatu; BEREZI
Modua	-tu beharrez	-tu beharrez	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	mendeko perpausean behar izan txertatu; BEREZI
Modua	-tu nahirik	-tu nahirik	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	mendeko perpausean nahi izan txertatu; BEREZI
Modua	-tu nahian	-tu nahian	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	mendeko perpausean nahi izan txertatu; BEREZI

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Modua	-tu gurarik	-tu gurarik	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	mendeko perpausean nahi izan txertatu; BEREZI
Modua	-tu ezinik	-tu ezinik	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	mendeko perpausean ezin izan txertatu; BEREZI
Modua	-tu ezinda	-tu ezinda	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	mendeko perpausean ezin izan txertatu; BEREZI
Modua	-tu ezinean	-tu ezinean	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	mendeko perpausean ezin izan txertatu; BEREZI
Modua	-tu ahala	-tu ahala	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	mendeko perpausaren aditzari -ten aspektua gehitu; BEREZI
Modua	-tu arau	-tu arau	Hala	Horrela	Era berean	Modu horretan	nagusiaren hasieran	mendeko-a_{jat}-na-gusia$_{jat}$	mendeko perpausaren aditzari -ten aspektua gehitu; BEREZI
Modua	-tzeko moduan	ø	ø	ø	ø	ø	ø	ø	EZ SINPLIFIKATU

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Modua	-tzeko gisan	ø	ø	ø	ø	ø	ø	ø	EZ SINPLIFIKATU
Modua	-tzeko eran	ø	ø	ø	ø	ø	ø	ø	EZ SINPLIFIKATU
Modua	-tzeko maneran	ø	ø	ø	ø	ø	ø	ø	EZ SINPLIFIKATU
Modua	-tzeko moldean	ø	ø	ø	ø	ø	ø	ø	EZ SINPLIFIKATU
Modua	-tu bezala	ø	ø	ø	ø	ø	ø	ø	EZ SINPLIFIKATU
Modua	-tzekotan	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Ondorioa	non... bait-	non... bait-	Ondorioz	Beraz	Hortaz	Honenbestez	mendekoaren hasieran	nagusia$_{jat}$-mendekoa$_{jat}$	kuantifikatzaile aldaketa; BEREZI
Ondorioa	ezen... bait-	ezen... bait-	Ondorioz	Beraz	Hortaz	Honenbestez	mendekoaren hasieran	nagusia$_{jat}$-mendekoa$_{jat}$	kuantifikatzaile aldaketa; BEREZI
Ondorioa	non... -en	non... -en	Ondorioz	Beraz	Hortaz	Honenbestez	mendekoaren hasieran	nagusia$_{jat}$-mendekoa$_{jat}$	kuantifikatzaile aldaketa; BEREZI
Ondorioa	ezen... -en	ezen... -en	Ondorioz	Beraz	Hortaz	Honenbestez	mendekoaren hasieran	nagusia$_{jat}$-mendekoa$_{jat}$	kuantifikatzaile aldaketa; BEREZI

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Helburua	subjuntiboa	subjuntiboa	nahi izan	gura izan	ø	ø	mendeko perpausean	nagusia$_{jat}$-mendekoa$_{jat}$	mendeko perpausaren aditza nominalizatu; aditz laguntzailea ezabatu; BEREZI
Helburua	-tzeko	-tzeko	nahi izan	gura izan	ø	ø	mendeko perpausean	nagusia$_{jat}$-mendekoa$_{jat}$	mendeko perpausaren aditza partizipioan jarri; BEREZI
Helburua	-tzekotz	-tzekotz	nahi izan	gura izan	ø	ø	mendeko perpausean	nagusia$_{jat}$-mendekoa$_{jat}$	mendeko perpausaren aditza partizipioan jarri; BEREZI
Helburua	- tzekotzat	- tzekotzat	nahi izan	gura izan	ø	ø	mendeko perpausean	nagusia$_{jat}$-mendekoa$_{jat}$	mendeko perpausaren aditza partizipioan jarri; BEREZI
Helburua	-tzearren	-tzearren	nahi izan	gura izan	ø	ø	mendeko perpausean	nagusia$_{jat}$-mendekoa$_{jat}$	mendeko perpausaren aditza partizipioan jarri; BEREZI
Helburua	- tzeagatik	- tzeagatik	nahi izan	gura izan	ø	ø	mendeko perpausean	nagusia$_{jat}$-mendekoa$_{jat}$	mendeko perpausaren aditza partizipioan jarri;BEREZI

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Helburua	-tze alde	-tze alde	nahi izan	gura izan	ø	ø	mendeko perpausean	nagusia$_{jat}$-mendekoa$_{jat}$	mendeko perpausaren aditza partizipioan jarri; BEREZI
Helburua	-tze aldera	-tze aldera	nahi izan	gura izan	ø	ø	mendeko perpausean	nagusia$_{jat}$-mendekoa$_{jat}$	mendeko perpausaren aditza partizipioan jarri; BEREZI
Helburua	-tzekotan	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Helburua	-tzeko asmotan	ø	ø	ø	ø	ø	ø	ø	EZ SINPLIFIKATU
Helburua	-tzeko intentziotan	ø	ø	ø	ø	ø	ø	ø	EZ SINPLIFIKATU
Helburua	-tzeko intentzioarekin	ø	ø	ø	ø	ø	ø	ø	EZ SINPLIFIKATU

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Baldintza	ba- (errealaorain)	ba-	Demagun	Suposa dezagun	Kasu horretan	ø	Txertatzeko eta txertatzeko2 mendekoan; txertatzeko3 nagusian	mendekoa_{jat}-nagusia$_{jat}$	baldin lokailua badago, hori ere ezabatu; BEREZI
Baldintza	ba- (errealalehen)	ba-	Demagun	Suposa dezagun	Kasu horretan	ø	Txertatzeko eta txertatzeko2 mendekoan; txertatzeko3 nagusian	mendekoa_{jat}-nagusia$_{jat}$	baldin lokailua badago, hori ere ezabatu; BEREZI
Baldintza	ba- (irreala)	ba-	Bestela	ø	ø	ø	nagusiaren hasieran	mendekoa_{jat}-nagusia$_{jat}$	polaritate aldaketa mendeko perpausean; baldin lokailua badago, hori ere ezabatu; BEREZI
Baldintza	non eta ez... -en	ø	ø	ø	ø	ø	ø	ø	Korrelazioak

(Jarraipena hurrengo orrialdean)

Mota	Egitura	Ezaba.	Txert.	Txert2	Txert3	Txert4	Txertatu non?	Hurrenk.	Oharrak
Baldintza	non ez... -en	ø	ø	ø	ø	ø	ø	ø	Korrelazioak
Baldintza	-tuz gero	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tuz geroz	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tuz geroztik	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tu ezkero	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tu ezkeroan	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tu ezkerok	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tu ezkerotino	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tu ezkeroz	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Baldintza	-tu ezkeroztik	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tu ezik	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tu ezean	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tzekotan	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tzekoz	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tzez gero	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tuenean	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa
Baldintza	-tzera	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa

(Jarraipena hurrengo orrialdean)

Mota	**Egitura**	**Ezaba.**	**Txert.**	**Txert2**	**Txert3**	**Txert4**	**Txertatu non?**	**Hurrenk.**	**Oharrak**
Baldintza	-tzerat	ø	ø	ø	ø	ø	ø	ø	Ordezkapen sintaktikoen sinplifikazioa

B.6 taula – Euskarazko egiturak sinplifikatzeko erregelak (perpaus adberbialak)

C

ETSC corpusa eskuz sinplifikatzean bete beharreko eragiketen zerrenda

Eranskin honetan ETSC corpusaren handitze-fasean testuak eskuz sinplifikatzean egin behar diren eragiketen zerrenda aurkeztuko dugu. Eragiketa horiek dira:

- Transformazioak batez ere sintaxian egin
- Perpaus koordinatuak banatu eta horietatik esaldi berriak sortu
- Perpaus adberbialak banatu, kontzesio- eta kausa-perpausak batez ere, eta horietatik esaldi berriak sortu
- Perpaus erlatiboak banatu eta horietatik esaldi berriak sortu
- Postposizio-sintagmak banatu eta horietatik esaldi edo sintagma berriak sortu
- Informazio osagarria eman (informazioa esplizitu egin, definizioak eman)
- Eliditutako argumentuak berreskuratu (koreferentzia ebatzi)
- Jatorrizko testuak dauden akatsak zuzendu

Tesi honen idazketa
2016ko maiatzaren 17an
bukatu zen.