EKAIA Ale berezia (2017). DOI: 10.1387/ekaia.17907

Euskarazko testu idatzien konplexutasunaren azterketa eta sinplifikazio automatikorako proposamena

(Analysis of Readability of Basque Complex Written Texts and the Proposal for their Automatic Simplification)

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza

Ixa Taldea, Euskal Herriko Unibertsitatea (UPV/EHU)

itziar.gonzalezd@ehu.eus

Jasoa: 2017-06-05

Onartua: 2017-07-12

Laburpena: *Gure gizartean egunero milioika testu sortzen dira, eta ikerketa honen helburua testu horiek ulerterrazagoak egitea da. Izan ere, horietariko asko konplexuak direnez, ez dira eskuragarriak arazo kognitiboak dituzten pertsonentzat edo atzerriko hizkuntzak ikasten ari direnentzat, besteak beste. Testu konplexuetatik informazioa erauztea ere ez da lan erraza Hizkuntzaren Prozesamendua egiten duten sistementzat. Arazo horiei aurre egiteko, tesi-lan honetan euskarazko testu idatzien konplexutasuna aztertu dugu eta, konplexutasun hori tratatzeko helburuarekin, testuen sinplifikazio automatikorako proposamena egin dugu.*

Hitz gakoak: konplexutasunaren analisia, testuen sinplifikazioa, euskara, hizkuntzaren prozesamendua.

Abstract: Millions of texts are produced every day in our society, and the aim of this research is to make them easier to understand. In fact, many of them are not accessible due to their complexity for people with cognitive disabilities or foreign language learners among others. Extracting information from complex texts is also difficult for Natural Language Processing applications. To overcome the problems text complexity causes, in this PhD. thesis we have analysed the readability of Basque texts and we have made a proposal to simplify the complex ones automatically.

Keywords: Readability Assessment, Text Simplification, Basque, Natural Language Processing.

1. SARRERA

Artikulu honetan "Euskarazko egitura sintaktiko konplexuen analisirako eta testuen sinplifikazio automatikorako proposamena / Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures" tesi-lana [1] aurkeztuko dugu. Lan hori Euskal Herriko Unibertsitateko Ixa ikerketa-taldean[1] garatu da, eta Hizkuntzalaritza eta Informatika alorrak uztartzen ditu.

Tesi-lan honek helburu hauek ditu: alde linguistikotik, testuen konplexutasuna aztertzea eta sinplifikazio-proposamenak egitea; eta alde konputazionaletik, sistema automatikoa diseinatzea, informazio linguistikoarekin hornitzea eta inplementatzea. Izan ere, egitura konplexuek arazoak sortzen dizkiete hizkuntzak ikasten ari diren pertsonei, arazo kognitiboak dituztenei edo alfabetizazio maila baxua dutenei, gutxi batzuk aipatzearren. Baina pertsonak ez dira arazo horiek jasaten dituzten bakarrak: informazioa eta testuak prozesatzen dituzten Hizkuntzaren Prozesamenduko (HP) tresnek zailtasunak dituzte esaldi konplexuak eta luzeak prozesatzean.

Egitura konplexuek sortzen dituzten arazoei aurre egiteko, testuen konplexutasunaren analisia (*Readability Assessment*) eta testuen sinplifikaziorako teknikak (*Text Simplification*) beharrezkoak dira. Lehenengoak testu bat konplexua den ala ez edo zein konplexutasun maila duen aztertzen du; bigarrenak, aldiz, testu konplexuak sinpleago bihurtzea du helburu, betiere jatorrizko esanahiari eutsiz. Bi ikerketa-lerro horiek irakaskuntzaren alorrean hasi baziren ere, egun HPan garrantzitsuak bihurtu dira, testuen konplexutasuna banan-banan aztertzea eta testu horiek eskuz sinplifikatzea ataza garestiak eta motelak direlako.

Aipatutako helburu horiek lortzeko, corpusak eta testuen analisi automatikorako oinarrizko tresnak erabili ditugu. Baliabide horietako batzuk jada eskuragarri zeuden, Euskararen Prozesamendurako Erreferentzia Corpusa (EPEC corpusa) [2] eta Ixa taldearen analisi-katea [3], esaterako. Beste batzuk, adibidez, Mugak [4] eta Aposizioak [5] izeneko tresnak, guk hobetu eta sortu ditugu.

Hurrengo ataletan, tesi-lan honetan erabilitako tresnak aurkeztuko ditugu (2. atala). Ondoren, testuen konplexutasunaren analisian (3. atala) eta testuen

[1] http://ixa.eus/ (atzitze-data: 2017-05-10)

sinplifikazioan (4. atala) egindako lanak azalduko ditugu. Tesi-lanaren ondorio eta etorkizuneko lan garrantzitsuenekin (5. atala) amaituko dugu.

2. TRESNEN EGOKITZAPENA

Testuen konplexutasuna automatikoki neurtzeko eta sinplifikazioa automatikoki gauzatu ahal izateko, testuak aldez aurretik analizatu behar ditugu. Analisi hori egiteko, HPko oinarrizko tresnak erabili ditugu.

Erabili dugun tresna multzorik garrantzitsuena Ixa taldearen analisi-katea [3] da. Kate horretan dauden tresnek testuak morfologikoki analizatu, lematizatu, morfosintaktikoki desanbiguatu eta zatitu (entitate-izenak, menderagailuak, postposizioak, sintagmak eta aditz-kateak ezagutu, eta funtzio sintaktikoak desanbiguatu) egiten dituzte.

Baina tresna horiek ez dira nahikoak konplexutasuna automatikoki aztertzeko eta sinplifikazioa automatikoki egiteko: perpausen mugak eta aposizioak detektatu behar ditugu. Horregatik, MuGa gramatika [6] hobetu dugu eta Aposizioak [5] izeneko tresna garatu dugu. MuGa gramatikak esaldien eta perpausen mugak identifikatzen ditu, eta Mugak [4] tresnak MuGa gramatikaren informazioan eta heuristikotan oinarrituta perpaus horiek banatzen ditu. Aposizioak tresnak, berriz, aposizioak identifikatzen, sailkatzen eta banatzen ditu. Bi tresna horiek ezagutza linguistikoa dute oinarrian.

3. KONPLEXUTASUNAREN ANALISIA

Tesi-lan honetan, konplexutasunaren analisia egiteko bi bide jorratu ditugu: alde batetik, EPEC[2] [2] eta Elhuyar[3] [7] corpusetan egitura sintaktiko konplexuak aztertu ditugu gure hurbilpena osatzeko; beste aldetik, testuak sinple ala konplexu bezala sailkatzen dituen sistema inplementatzeko, esanguratsuak izan daitezkeen ezaugarri linguistikoak aztertu ditugu ikasketa automatikoko teknikak erabilita.

[2] EPEC corpusa euskara batuan idatzitako 300.000 hitzeko bilduma da, hainbat maila linguistikotan etiketatuta dago, eta euskara automatikoki prozesatzeko erreferentziazko corpusa da.
[3] Elhuyar corpusa izen bereko aldizkariko erreportajeez eta albisteez osatuta dago eta testuen konplexutasunaren analisia egiten duen sistema entrenatzeko bildu da.

3.1 Corpus-azterketan oinarritutako konplexutasunaren analisia: testuen sinplifikazio automatikoari bideratutako hurbilpena

Corpus-azterketa egiteko, ingeleserako eta Brasilgo portugeserako egindako Siddharthan-en [8] eta Specia eta besteren [9] lanetan konplexutzat hartutako fenomenoak EPEC eta Elhuyar corpusetatik erauzi ditugu, eta horien azterketa eta sailkapena eginez, euskaraz egitura konplexuak definitzeko irizpideak finkatu eta sinplifikazio-proposamenak egin ditugu. Horretaz gain, EPEC corpusean fenomeno horien maiztasunak eta, perpaus adberbialen kasuan, kokapena ere aztertu ditugu sinplifikazio-proposamenak egiteko.

Zehazki, hauexek dira azterketa honetan konplexutzat hartu ditugun fenomenoak: perpaus koordinatuak, mendeko perpausak, aposizio-sintagmak, informazio biografikoa duten egitura parentetikoak eta adierazpenak adierazten dituzten postposizio-sintagmak dira.

Aipatutako fenomeno konplexuen sinplifikazio-proposamenak ere corpus-azterketan oinarrituta egin ditugu. Proposamen horiek gure hurbilpenean egin daitezke: i) egitura-aldaketarik[4] gabe (perpaus adberbial ez-jokatuentzat soilik) eta ii) egitura-aldaketekin. Jarraian, proposamen horiek laburbilduko ditugu.

Ordezkapen sintaktikoen sinplifikazioa: egitura-aldaketarik gabe

Ordezkapen sintaktikoen sinplifikazioa izeneko sinplifikazio motan, ez da egitura-aldaketarik egiten eta maiztasun gutxiko egiturak maiztasun altuagoa duten baliokideekin ordezkatzen dira. Hori azaleko ordezkapen sintaktikoa eragiketaren bidez gauzatzen da. Horrela, (1) adibidean EPEC corpusean maiztasun txikia duen *–tu beharrean* egitura maiztasun handiagoa eta esanahiaren aldetik baliokidea den *–tu ordez* egiturarekin ordezkatu da. Aldaketa horiek adibidean azpimarratu ditugu.

(1) **Jatorrizkoa:** *Zabortegia egin beharrean, turismoa erakar dezaketen proiektuak bultzatu beharko genituzke, beharbada, Nobel Sarien arrakasta aprobetxatuz.*

a. **Sinplifikatutakoa1:** *Zabortegia egin ordez, turismoa erakar dezaketen proiektuak bultzatu beharko genituzke, beharbada, Nobel Sarien arrakasta aprobetxatuz.*

[4] *Egitura-aldaketa* terminoarekin adierazi nahi dugu jatorrizko esaldian mendekotasun-erlazioak kentzen direla.

Sinplifikazio mota horren bitartez moldatutako testuak azaleko sinplifikazio sintaktiko (ASS) sinplifikazio mailara egokitzen dira. Maila hori zuzenduta dago sintaxia menderatzen duten baina egitura guztiak ezagutzen ez dituzten pertsonei (euskara maila aurreratua, B2tik aurrera) eta entrenamendu-corpusetan maiztasun gutxiko egiturak izan ez dituzten HPko tresnei.

Sinplifikazio sintaktikoa: egitura-aldaketekin

Sinplifikazio sintaktikoko sinplifikazio-proposamenenetan egitura-aldaketak egiten dira. Hurrengo puntuetan prozedura hori azalduko dugu, bakoitzari dagokion eragiketa parentesi artean adieraziz:

a) Esaldian dauden perpausak banatu (banaketa)

b) Perpausetatik esaldi sinplifikatuak berreraiki erlazio-markak ezabatuta eta ezabatutako elementuen esanahia berreskuratzeko txertatze-elementuak txertatuta (esaldien berreraikitzea, ezabatzea eta txertatzea)

c) Sortutako esaldi berriak testuan ordenatu (esaldien ordenatzea)

d) Esaldiak (ortografikoki eta ortotipografikoki) zuzenak diren egiaztatu (esaldien zuzenketa eta egokitzapena)

Prozesu horren emaitza (2) adibideko esaldietan ikus daiteke:

(2) **Jatorrizkoa:** *Edurnezuri printzearekin <u>ezkondu zenean</u>, zazpi ipotxek edateari eman zioten.*
   a. **Sinplifikatutakoa1:** *Edurnezuri printzearekin <u>ezkondu zen</u>.*
   b. **Sinplifikatutakoa2:** *<u>Orduan</u>, zazpi ipotxek edateari eman zioten..*

Sinplifikazio sintaktikoan testuak bi mailatara egokitzen dira: sinplifikazio naturala (SN) eta sinplifikazio absolutua (SA). SNa tarteko euskara-maila dutenei (erdi-maila, B1-B2) eta esaldi laburrak hobeto prozesatzen dituzten tresnei zuzenduta dago; SA, berriz, euskara ikasten hasi direnei (hastapen-maila, A1-A2) eta oso esaldi laburretan informazioa erauzi behar duten tresnei.

Laburbilduz, helburu-taldeek izan ditzaketen beharrei erantzunez sortutako hiru sinplifikazio maila horietara egokitutako esaldi bat erakusten dugu (3) adibidean.

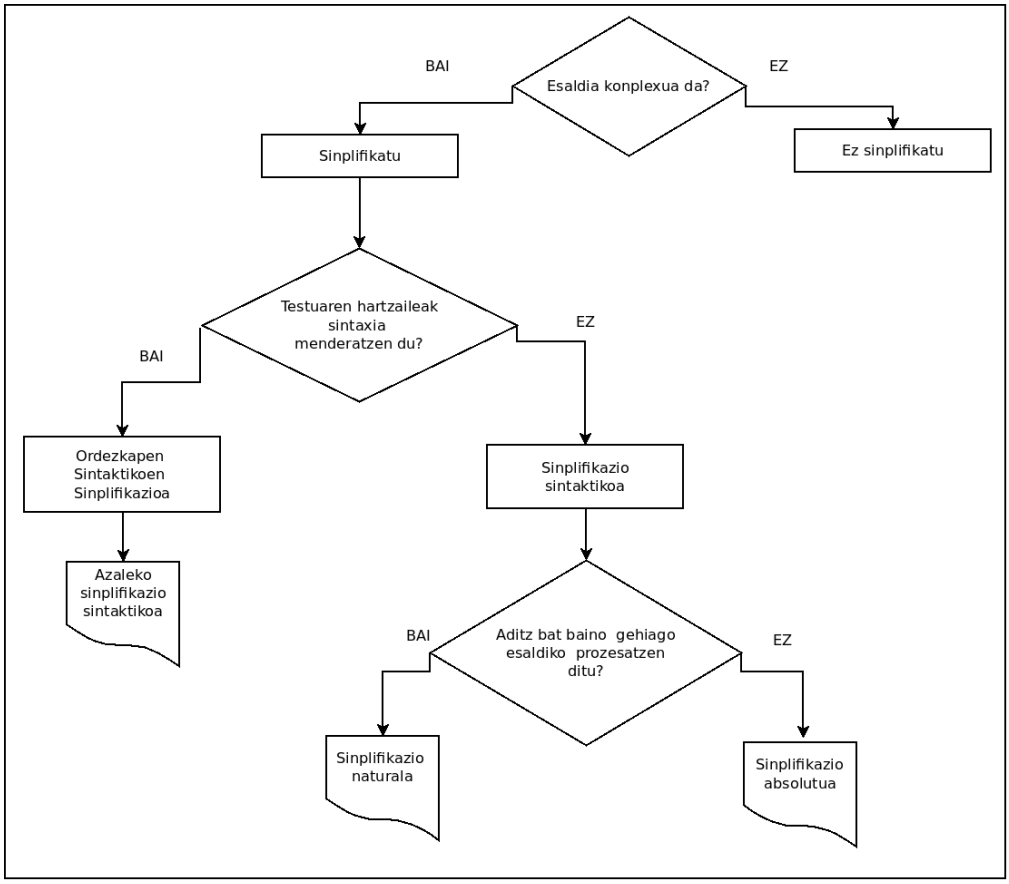
(3) **Jatorrizkoa:** *1991 eta 1993an errepideko Munduko Txapelketak eta bi Tour irabazi ondoren, mendian gora aise igotzearren pisua galtzen hasi zen, eta 1994. urtean 48 kiloko infernura jaitsi zen, anorexiara.*

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

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

SA: *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 mota eta maila aukeratzeko, 1. irudian aurkezten dugun sinplifikazio-erabakien algoritmoa erabili dugu. Algoritmo horren bitartez erabakitzen da sinplifikazioa egin behar den edo ez; eta egin behar bada, zein motatakoa (ordezkapen sintaktikoen sinplifikazioa edo sinplifikazio sintaktikoa) eta zein mailatakoa (azaleko sinplifikazio sintaktikoa, sinplifikazio naturala edo sinplifikazio absolutua) gauzatu behar den.



1. irudia. Sinplifikazio-erabakien algoritmoa.

3.2 Konplexutasunaren analisi automatikoa

Konplexutasunaren analisi automatikoan, gure helburua da testuen konplexutasuna adierazten duten ezaugarriak ezagutzea, testuak sinplifikatu behar diren ala ez jakiteko. Alegia, testua konplexua edo sinplea den jakin nahi dugu sinplifikatu aurretik. Horretarako, sei maila linguistikotan banatzen diren 94 ezaugarri (ratio) inplementatu ditugu. Horiek kontuan hartuta, ikasketa automatikoko esperimentuak egin ditugu sailkatzaileak entrenatzeko. Ezaugarri linguistiko horietatik estatistikoki esanguratsuenak diren lehen hamarrak 1. taulan zerrendatu ditugu. Beste ezaugarri batzuk dira, adibidez, esaldien luzera, mendeko perpausen ratioa edo orainaldiko perpausen ratioa[5]. Entrenamendurako erabili ditugun corpusak Elhuyar[6] eta Zernola[7] [7] izan dira.

1. taula. 10 ezaugarri esanguratsuenak.

Mota	Ezaugarria (ratioa)	Esanguratsutasuna
Lexikala	Izen berezien eta izen arrunten arteko ratioa	0,2744
Morfosintaxikoa	Aposizio-sintagmen eta izen-sintagmen arteko ratioa	0,2529
Morfosintaxikoa	Aposizio-sintagmen eta sintagmen arteko ratioa	0,2529
Lexikala	Entitateen eta izen arrunten arteko ratioa	0,2436
Lexikala	Behin bakarrik agertzen diren lemen eta lema guztien arteko ratioa	0,2394
Lexikala	Siglen eta hitz guztien arteko ratioa	0,2376
Lexikala	Aditz faktitiboen eta aditz guztien arteko ratioa	0,2099
Sintaktikoa	Modu/denbora-perpausen eta mendeko perpaus guztien arteko ratioa	0,2056
Morfologikoa	Destinatiboaren eta kasu-marka guztien arteko ratioa	0,1968
Pragmatikoa	Azalpenezko lokailuen eta lokailu guztien arteko ratioa	0,1957

[5] Ratioak lortzeko bi kopururen arteko zatiketa egin dugu. Adibidez, izen berezien eta izen arrunten arteko ratioa (r) lortzeko, izen berezien kopurua (x) eta izen arruntena (y) zatitu ditugu (r = x / y).
[6] Elhuyar corpusa testu konplexuen eredutzat hartu dugu.
[7] Zernola corpusak izen bereko webgunetik erauzitako artikuluak biltzen ditu, eta, 8-13 urte bitarteko haurrentzat bideratuta izanik, testu sinpleen eredu gisa hartu dugu.

Zehazki, ikasketa automatikoko sailkatzailea aukeratzeko, 3 esperimentu egin ditugu (2. taula): lehenengoan, ezaugarri linguistiko guztiak batera aplikatu ditugu eta asmatze-tasa % 89,50 izan da; bigarrenean, ezaugarriak mailaren arabera aplikatu ditugu, alegia, orokorrak bere aldetik, lexikalak bere aldetik.... eta emaitzarik onena ezaugarri lexikalekin lortu dugu (% 90,75eko asmatze-tasa); eta, azkenik, hirugarrenean, ezaugarriak multzokatu ditugu eta emaitzarik onena ezaugarri lexikalek, morfologikoek, morfosintaktikoek eta sintaktikoek osatutako multzoarekin lortu dugu (% 93,50eko asmatze-tasa). Hiru esperimentuetan sailkatzaile onena euskarri bektoredun makinak (SVM) izan dira. Esperimentu horiek guztiak WEKA ikasketa automatikorako tresnarekin [10] eta geruzako balidazio gurutzatua erabiliz egin ditugu.

2. taula. Esperimentuen laburpena.

Esperimentua	Ezaugarri multzoa	Asmatze-tasa
Ezaugarri linguistiko guztiak	Guztiak	89,50
Ezaugarri linguistikoak mailaren arabera	Lexikalak	90,75
Ezaugarri linguistikoak multzokatuta	Lexikalek, morfologikoek, morfosintaktikoek eta sintaktikoek osatutako multzoa	93,50

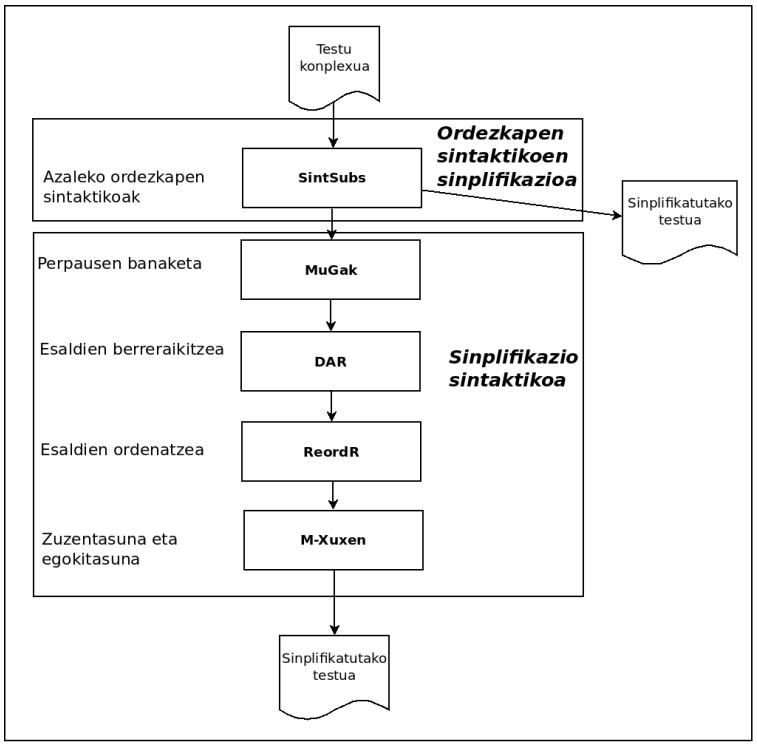
Prozesu honen emaitza ErreXail [7] izeneko sistema izan da.

4. TESTUEN SINPLIFIKAZIOA

Testuen sinplifikazioa tesi-lan honetan bi ikuspuntutatik aztertu dugu: batetik, testuak automatikoki sinplifikatuko dituen sistema diseinatu dugu (testuen sinplifikazio automatikoa) eta bestetik, guk proposatutako hurbilpena konparatzeko, eskuz sinplifikatutako testuen corpusa osatu eta analizatu dugu.

4.1 Testuen sinplifikazio automatikoa

Testuak automatikoki sinplifikatzeko EuTS sistema (2. irudia) diseinatu dugu. EuTS corpus-azterketan hartutako erabakietan oinarritzen da, eta erabaki horiek bost modulutan aplikatzen ditu.



2. irudia. EuTS sistemaren arkitektura.

Lehenengo moduluan, SintSubs izenekoan, ordezkapen sintaktikoen sinplifikazioa egiten du ordezkapen sintaktikoak eragiketaren bitartez. Modulu hori guztiz inplementatuta eta ebaluatuta dago eta, horretarako, *EGLU datu-multzoa* sortu dugu [11]. Datu-multzo hori *Euskal Gramatika Lehen Urratsak* (EGLU) gramatikako perpaus jokatugabeen liburukitik [12] bildutako esaldiekin osatu dugu.

SinSubs moduluak ematen duen emaitza (4) adibidean ikus daiteke: maiztasun gutxi duen helburuzko *–tzearren* egitura maiztasun handiagoa duen helburuzko *–tzeko* egiturarekin ordezkatu da.

(4) **Jatorrizkoa:** *Abuztuaren amaieran beste goi bilera bat egitea aztertzen ari dira Israel eta PAN Palestinako Aginte Nazionala, Ekialde Erdiko bake prozesua suspertzearren.*
   a. **Sinplifikatutakoa1:** *Abuztuaren amaieran beste goi bilera bat egitea aztertzen ari dira Israel eta PAN Palestinako Aginte Nazionala, Ekialde Erdiko bake prozesua suspertzeko.*

Hurrengo lau moduluetan EuTSek sinplifikazio sintaktikoa egiten du [13]. Lau modulu horiek Mugak, DAR, ReordR eta M-Xuxen dira, eta banaketa, esaldien berreraikitzea, esaldien ordenatzea eta esaldien zuzentasuna eta egokitzapena eragiketak aplikatzen dituzte hurrenez hurren. (5) adibidean modulu bakoitzak ematen duen irteera ikus daiteke kontzesio-perpaus jokatu batean gauzatuta.

(5) **Jatorrizkoa:**. *Asperren kasua emeki-emeki aitzinatu bada ere, Sa Pintoren etorkizuna fite argituko da.*
   a. **Mugak (banaketa):** [*Asperren kasua emeki-emeki aitzinatu bada ere,*][*Sa Pintoren etorkizuna fite argituko da*.]
   b. **DAR (esaldien berreraikitzea):**
      i. Ezabatzea: [*Asperren kasua emeki-emeki aitzinatu da*]
      ii. Txertatzea: [*Hala ere, Sa Pintoren etorkizuna fite argituko da*.]
   c. **ReordR (esaldien ordenatzea):** [*Asperren kasua emeki-emeki aitzinatu da*] [*Hala ere, Sa Pintoren etorkizuna fite argituko da*.]
   d. **M-Xuxen (esaldien zuzentasuna eta egokitasuna):** *Asperren kasua emeki-emeki aitzinatu da*. *Hala ere, Sa Pintoren etorkizuna fite argituko da*.

Modulu horietatik Mugak guztiz inplementatuta dago, eta DAR eta ReordR partzialki. Dena den, eragiketak aplikatu ahal izateko beharrezkoa den informazio guztia definitu dugu.

Kasu-azterketa gisa, EuTSen arkitekturari jarraituz, Biografix izeneko tresna eleaniztuna garatu dugu [14]. Biografixek informazio biografikoa duten egitura parentetikoetatik esaldi sinpleak sortzen ditu 8 hizkuntzatan patroietan oinarrituta. Horretarako, Wikipedia datu-multzoa [14] osatu dugu Wikipedia entziklopediatik[8] jasotako esaldiekin. Biografixek euskaraz nola jokatzen duen (6) adibidean ikus daiteke.

(6) **Jatorrizkoa:***Ernest Rutherford, Nelsongo lehenengo baroia, (Brightwater, Zeelanda Berria, 1871ko abuztuaren 30a – Cambridge, Ingalaterra,1937ko urriaren 19a) fisika nuklearraren aita izan zen.*
   a. **Sinplifikatutakoa1:** *Ernest Rutherford, Nelsongo lehenengo baroia, fisika nuklearraren aita izan zen.*
   b. **Sinplifikatutakoa2:** *Ernest Rutherford 1871ko abuztuaren 30ean, Brightwateren jaio zen.*
   c. **Sinplifikatutakoa3:** *Brightwater Zeelanda Berrian dago.*
   d. **Sinplifikatutakoa4:** *Ernest Rutherford 1937ko urriaren 19ean Cambridgen hil zen.*
   e. **Sinplifikatutakoa5:** *Cambridge Ingalaterran dago.*

Biografixen kodea[9] eskuragarri dago, euskarazko zein beste hizkuntzetarako garatu diren bertsioak hobetu edo moldatu nahi izanez gero.

4.2 Euskarazko testu sinplifikatuen corpusa (ETSC)

Sinplifikazioa gauzatzeko eta gure erabakiak eta hurbilpen ezberdinak kontrastatzeko, corpus bat osatu dugu jatorrizko 227 esaldirekin. Esaldi horiek Elhuyar

[8] Euskal Wikipedia https://eu.wikipedia.org/wiki/Azala (atzitze-data: 2017-03-27)
[9] http://ixa.si.ehu.es/node/4482?language=eu (atzitze-data: 2017-05-16)

corpusetik erauzi ditugu eta bi hizkuntzalariri sinplifikatzeko eskatu diegu. Hizkuntzalari bati gidalerro eta irizpide orokorrak eman dizkiogu testuak hurbilpen estrukturalaren aldetik sinplifika ditzan, eta beste hizkuntzalariari bere eskarmentuaren araberako sinplifikazioa egin dezan eskatu diogu; alegia, hurbilpen intuitiboaren aldetik sinplifika ditzan.

Egindako sinplifikazioak aztertzeko eta konparatzeko, 8 makroeragiketa biltzen dituen etiketatze-eskema osatu dugu. Makroeragiketa horiek dira ezabatzea, bateratzea, banaketa, transformazioa, txertaketa, hurrenkera-aldaketa, eragiketarik eza eta bestelakoak. 3. taulan azaltzen ditugu labur.

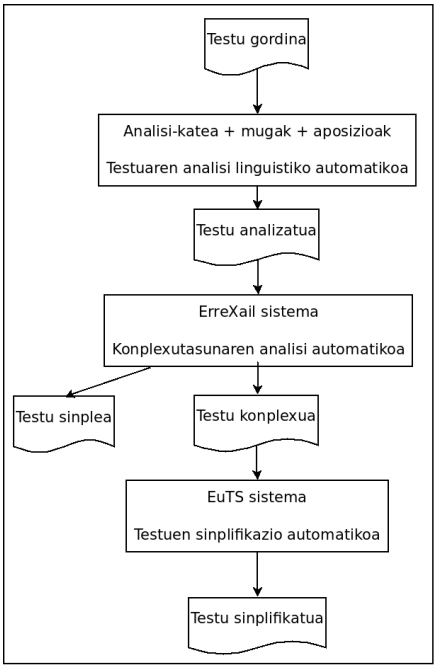
3. taula. Etiketatze-eskemako makroeragiketen azalpena.

Makroeragiketa	Azalpena
Ezabatzea	Kasu-markak, hitzak, sintagmak, perpausak edo esaldiak ezabatzea
Bateratzea	Perpaus/esaldi bat baino gehiagotatik perpaus/esaldi bat sortzea
Banaketa	Sintagmak, perpausak edo esaldiak banatzea
Transformazioa	Jatorrizko hitzak, sintagmak, perpausak edo esaldiak eraldatzea
Txertaketa	Elementu berriak (hitzak, sintagmak, perpausak edo esaldiak) txertatzea
Hurrenkera-aldaketa	Hitzen, sintagmen, perpausen edo esaldien hurrenkera aldatzea
Eragiketarik eza	Eragiketarik ez egitea
Bestelakoak	Bestelakoak edo eragiketen konbinazioa

Emaitzei dagokienez, bi hurbilpenetan makroeragiketarik erabiliena transformazioa izan da, eta transformazio motarik erabiliena sintaktikoa. Perpausen banaketari dagokionez, biek banatu dituzte koordinazioak eta perpaus adberbialak gehien. Emaitza horiek bat egiten dute gure proposamenarekin sinplifikazioa sintaxian egitea erabaki dugulako eta banaketak gure sinplifikazio-prozesuaren oinarria direlako. Gainontzeko emaitzak gure etorkizuneko lanarekin bat datoz. Beraz, testuak sinplifikatzean, honako hauek kontuan izan behar direla ondorioztatu dugu: sintaxi mailako sinplifikazioa egitea eta nahitaezkoa ez den informazioa gehitzea, elidi tutako subjektuak, objektuak eta abar berreskuratuz.

5. ONDORIOAK ETA ETORKIZUNEKO LANAK

Tesi-lan honetan, beraz, euskarazko konplexutasun sintaktikoa analizatzeko eta egitura konplexuak sinplifikatzeko lehen proposamena egin dugu. Horretarako, atzerriko hizkuntzetan egin diren lanak eta euskarazko corpusen azterketak izan ditugu oinarri euskarazko testuak konplexutasunaren arabera sailkatzen dituen sistema inplementatzeko eta, konplexuak izanez gero, sinplifikatuko dituen sistema diseinatzeko. Prozesu guztia 3. irudian laburbildu dugu.



3. irudia. Testuen konplexutasuna neurtzeko eta sinplifikatzeko prozesua.

Etorkizunerako, ErreXail sistemari ezaugarri gehiago gehitu nahi dizkiogu, rol semantikoekin lotutakoak, esaterako, eta konplexutasun maila gehiago bereizteko entrenatu nahi dugu. Testuen sinplifikazio automatikoari dagokionez, EuTS sistemaren inplementazioa bukatu eta ebaluatu nahi dugu. Horren ondorioz, ziurrenik azterketa linguistikoa sakondu eta erregelak birfindu beharko ditugu. Horretaz gain, sisteman lehentasunak ezartzeko balia ditzakegu ETSC corpusean egin eta egiteko ditugun analisiak.

ESKER ONAK

Tesi-lan hau Eusko Jaurlaritzak doktoreak ez diren ikertzaileak prestatzeko Doktoratu Aurreko Programako laguntza bati esker egin da.

6. BIBLIOGRAFIA

[1] GONZALEZ-DIOS I. 2016. *Euskarazko egitura sintaktiko konplexuen analisirako eta testuen sinplifikazio automatikorako proposamena / Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures.* Euskal Herriko Unibertsitatea (UPV/EHU).

[2] 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. 2006. «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.

[3] 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.

[4] ARANZABE M.J., DÍAZ DE ILARRAZA A. eta GONZALEZ-DIOS I. 2013. «Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque». *Procesamiento de Lenguaje Natural*, **50,** 61-68.

[5] GONZALEZ-DIOS I., ARANZABE M.J., DÍAZ DE ILARRAZA A. eta SORALUZE A. 2013. «Detecting Apposition for Text Simplification in Basque». *International Conference on Intelligent Text Processing and Computational Linguistics*, 513-524.

[6] ADURIZ I, ARRIETA B., ARRIOLA J.M., DÍAZ DE ILARRAZA A., IZAGIRRE E. eta ONDARRA A. 2006. *Muga Gramatikaren Optimizazioa.* Barne-txostena, UPV/EHU/LSI/TR 9-2006.

[7] GONZALEZ-DIOS I., ARANZABE M.J., DÍAZ DE ILARRAZA A. eta SALABERRI H. 2014. «Simple or Complex? Assessing the Readability of Basque Texts». *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 334-344.

[8] SIDDHARTHAN A. 2002. «An Architecture for a Text Simplification System». *Proceedings of the Language Engineering Conference*, 64-71.

[9] SPECIA L., ALUÍSIO eta PARDO T.A. 2008. *Manual de Simplificação Sintática para o Português.* Barne-txostena NILC-TR-08-06.

[10] HALL M., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN P. eta WITTEN I.H. 2009. «The WEKA Data Mining Software: an Update». *ACM SIGKDD Explorations Newsletter*, **11**(1),10-18.

[11] GONZALEZ-DIOS I., ARANZABE M.J. eta DÍAZ DE ILARRAZA A. 2015. «Simplifying Basque Texts: the Shallow Syntactic Substitution Simplification». Proceedings the 7th Language & Technology Conference, 450-454.

[12] EUSKALTZAINDIA. 2011. *Euskal Gramatika Lehen Urratsak: VII, (Perpaus jokatugabeak: denborazkoak, kausazkoak eta helburuzkoak, baldintzazkoak, kontzesiozkoak, moduzkoak, erlatiboak eta osagarriak).* Euskaltzaindia, Bilbo.

[13] ARANZABE M.J., DÍAZ DE ILARRAZA A. eta GONZALEZ-DIOS I. 2012. «First Approach to Automatic Text Simplification in Basque». *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, 1-8.

[14] GONZALEZ-DIOS I., ARANZABE M.J. eta DÍAZ DE ILARRAZA A. 2014. «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.