I. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2015eko maiatzaren 13, 14 eta 15
Durango, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

GIZA ZIENTZIAK

Euskararen Sorkuntza Automatikoa: lehen urratsak

M. Agirrezabal, I. Gonzalez-Dios eta *I. Lopez-Gazpio*

15-23 or.
https://dx.doi.org/10.26876/ikergazte.i.01

ANTOLATZAILEA: udako euskal unibertsitatea
BABESLEAK: Eusko Jaurlaritza Gobierno Vasco; BFA DFB Bizkaiko Foru Aldundia Diputación Foral de Bizkaia; eman ta zabal zazu UPV EHU
LAGUNTZAILEAK: Deusto Universidad de Deusto Deustuko Unibertsitatea; Euskal Herria Udalbiltza; Mondragon Unibertsitatea; upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa

Euskararen Sorkuntza Automatikoa: lehen urratsak

Agirrezabal, M. eta **Gonzalez-Dios, I.** eta **Lopez-Gazpio, I.**

Ixa Taldea. Lengoaia eta sistema informatikoak saila. Informatika fakultatea (UPV/EHU).
314 bulegoa
Manuel Lardizabal, 1
20018 Donostia

Laburpena

Lan honetan euskararen sorkuntza automatikoan eman diren lehen urratsak azaltzen ditugu. Urrats horiek azaltzeko, hiru ikerketa-lerrotan egin diren lanak aurkezten ditugu: i) poesiaren sorkuntzan, ii) testuen sinplifikazioan eta iii) galderen sorkuntzan. Atzerriko hizkuntzetan egin den lana azaldu ondoren, ikerketa-lerro horietan hizkuntzaren sorkuntza nola erabiltzen den banan-banan deskribatzen dugu adibideekin batera. Azkenik, etorkizunerako planteatu ditugun erronkak eta soluzio posibleak zerrendatzen ditugu.

Hitz gakoak: Hizkuntzaren Sorkuntza Automatikoa, Poesia Sorkuntza, Testuen Sinplifikazioa, Galderen Sorkuntza, Euskara

Abstract

In this paper we present the first steps towards Natural Language Generation in Basque. To explain this steps, we present the works carried out in three NLP research areas: i) Poetry Generation, ii) Text Simplification and iii) Question Generation. After showing what has been done for other languages, we present how we tackle the problem of Natural Language Generation in Basque. Finnaly, we discuss the challenges we face and outline the future work.

Keywords: Natural Language Generation, Poetry Generation, Text Simplification, Question Generation, Basque

1 Sarrera eta motibazioa

Hizkuntzaren Prozesamendua[1] (HP), ingelesez *Natural Language Proccesing (NLP)*, informatikaren, adimen-artifizialaren eta hizkuntzalaritzaren diziplinak konbinatzen dituen alorra da. HPan honako alor hauek aurkitzen dira besteak beste: analisi sintaktikoa, hitzen adiera-desanbiguazioa, sentimenduen analisia, korreferentziaren ebazpena, itzulpen automatikoa, hizketaren tratamendua, testuen laburpen automatikoa, eta informazioaren berreskurapena (Jurafsky eta Martin, 2000).

HPren helburu zein erronka nagusia hizkuntza ulertzeko gai izango diren sistema konplexuak eraikitzea da. Aitzitik, zeregin hau hasiera batean pentsa daitekeena baino askoz ere neketsuagoa da, makinek inguratzen gaituen munduaren noziorik ez dutelako. Gizakiok jaio-berritik ikasten ditugun erregela sinpleenak ere ezagutu behar dituzte, hizkuntza interpretatu zein sortzeko gai izango badira; eta, ezagutza hori makinei bereganatzea ez da lan xamurra. HPak eskaintzen dituen ikerketa-lerroen artean Hizkuntzaren Sorkuntzari (HS), ingelesez, *Natural Language Generation (NLG)*, helduko diogu artikulu honetan, azken urteetan pil-pilean dagoen alorra baita (Reiter eta Dale, 2000; Stent eta Bangalore, 2014).

Hizkuntzaren sorkuntza automatikoa HPko ataza bat da. Honen helburua hizkuntza sortzea da, gizakiok ulertzeko modukoa. Sorkuntza egiteko abiapuntuak errepresentazio abstraktuak (*concept-to-text generation*, *data-to-text generation*) edo testu arruntak (*text-to-text generation*) dira eta horien emaitza testu arruntak dira. HS sistemen arkitekturaz at emaitzaren prozesaketa geratzen da; adibidez, bertso-

[1]Prozesamendu automatikoaz arituko gara artikulu honetan, hau da, ordenagailuek edo makinek egiten duten prozesamendu automatikoaz eta ez garunaren prozesamenduaz.

sorkuntza sistema baten kasuan, bertsoa kantatzea edo testu sinplifikatuen eta galderen kasuan, testuaren formatuaren zehaztapena da HStik kanpo geratzen dena. Gure helburua euskara automatikoki sortzea da eta lan horretan Euskal Herriko Unibertsitateko Ixa ikerketa-taldean[2] ari gara.

HSk eskaintzen duen teknologia HPko hainbat aplikazioetan erabiltzen da: laburpenen sorkuntzan, testuen sinplifikazioan, galderen sorkuntzan, eta literatura sorkuntzan, esaterako. Oro har, lan gehienak ingeleserako egin dira, baina azken urteetan gero eta gehiago dira beste hizkuntzentzat egiten diren lanak.

Artikulu honetan HSan euskaraz egin diren lanak jasoko ditugu. Alde batetik, bertsoen sorkuntza automatikoan egiten ari garen lanak (Agirrezabal *et al.*, 2013), testuen sinplifikazio automatikokoak (Aranzabe *et al.*, 2012; Gonzalez-Dios, 2014) eta amaitzeko galderen sorkuntzaren ingurukoak (Lopez-Gazpio, 2013; Aldabe *et al.*, 2013).

Sarrera honen ondoren, 2. atalean poesiaren sorkuntza, testuen sinplifikazioa eta galderen sorkuntza zer diren azalduko dugu eta 3. atalean ikerketa-lerro horietan euskaraz egiten ari garena azalduko dugu. 4. atalean ondorioak jasoko ditugu eta 5. atalean etorkizunerako planteatzen dugun lana adieraziko dugu.

2 Arloko egoera eta ikerketaren helburuak

Argi dago hizkuntzaren sorkuntza automatikoak berebiziko garrantzia duela hizkuntzaren prozesamenduan, eta horren adibide dira atal honetan aurkezten ditugun lanak edo baliabideak.

HS sistema baten helburua makinarentzat ulergarria den datu-multzo bat hizkuntza arruntera bihurtzea da. Hauen garapenerako hainbat egitura jarrai daitezke, Reiter eta Dale autoreen (2000) liburuan aipatutakoa, adibidez. Bertan, modu orokorrean azaltzen dute zer izan behar den kontuan HS sistema bat garatzeko eta arkitektura bat proposatzen dute.

HS sistema aipagarri bat *MULTIGEN* sistema da (Barzilay *et al.*, 2001), non laburpen automatikoak sortzen dituzten hainbat dokumentutatik abiatuta. Adibidez, gertakari bera aipatzen duten hainbat albistetan oinarrituta, albiste bakarra sortzen dute. Bestalde, eguraldi iragarpenak edo antzekoak testugisa sortzeko sistemak ere garatu dira (Goldberg *et al.*, 1994; Coch, 1998; Turner *et al.*, 2006). Azken bi sistemek datu-baseetatik erauzitako informazioa hizkuntza arruntean jartzea dute helburu. Hurrengo lerroetan poesiaren sorkuntzan, testuen sinplifikazioan eta galderen sorkuntzan egin diren lan esanguratsuenak aipatuko ditugu.

Poesiaren sorkuntza HPko helburu utopiko bat da. Konputagailuek poesia sortzeko balizko aukera Alan Turing matematikariaren 1950 urteko artikuluan aipatzen da. Azken urteotan, baina, sistema kreatiboek garrantzia hartu dute (Yan *et al.*, 2013; Zhang eta Lapata, 2014; Toivanen *et al.*, 2014; Gervás; Oliveira eta Cardoso, 2015), baita hauetako batzuk biltzen dituen liburuak ere (Besold *et al.*, 2014). Oliveira-k (2012) poesia-sortzaile moldagarri bat aurkeztu du, PoeTryMe izenekoa, poesia portugesez sortzen duena, eta ondoren egindako lan batean poesia-sortzaile hori gaztelaniarako moldatu dute (Oliveira *et al.*, 2014).

Testuen sinplifikazio automatikoa (TS) lehen aldiz 1996an aipatu zen eta, batez ere, 2008tik aurrera eztanda handia izan duen HPko ikerketa-lerroa da (Gonzalez-Dios *et al.*, 2013; Shardlow, 2014; Siddharthan, 2014). Bere helburua da testuen ulermena areagotzea testu horren sintaxia eta lexikoa ezagunagoa eginez, beti ere jatorrizko testuaren edukiari eta esanahiari eutsiz. HPn maiz gertatzen den bezala, hasierako lanak ingeleserako egin ziren arren azken urteetako eztandaren ondorioz TSa japonierara, Brasilgo portugesera, suedierara, arabierara, gaztelaniara, frantsesera, euskarara, italierara eta bulgarierara zabaldu da.

Testu sinplifikatuak onuragarriak dira bai pertsonentzat baita HPko aplikazio aurreratuentzat ere. Pertsonen artean nabarmendu ditzakegu besteak beste atzerriko hizkuntzak ikasten ari direnak, arazo kognitiboak dituztenak, entzumen arazoak dituztenak, dislexikoak, adineko pertsonak. HPko tresnek, aldiz, testu errazak eta esaldi laburrak dituztenak prozesatzean emaitza hobeak lortzen dituzte eta, beraz, testu sinplifikatuak ere egokiak dira itzultzaile automatiko, galdera-erantzun sistema eta bilatzaileentzat, bakar batzuk aipatzearren.

TSko sistemak gizakiek idatzitako erregeletan oinarritutakoak eta estatistikan oinarritutakoak dira.

[2] `http://ixa.eus/Ixa`

Erregeletan oinarritutako sistema batzuk irakurketa errazaren gomendioak jarraitzen dituzte eta besteak hizkuntzalariek idatzitako erregeletan oinarritzen dira. Sistema estatistikoek erregelak corpusetatik ikasten dituzte, ingeleseko Wikipedia eta *simple Wikipedia* erabiliz, esaterako. Beste sistema batzuk testuen sinplifikazioa itzulpen automatikoa bezala ulertzen dute, hizkuntza konplexutik hizkuntza sinplerako itzulpen bezala hain zuzen ere.

HSko teknologia hezkuntzaren edo pedagogiaren alorrean ere oso baliagarria da, esaterako, testuen-ulermena lantzeko aukera ematen duelako. Testu-ulermena bizi garen heinean erabilgarri dugun gaitasuna da, momentuero garatuz doana eta irakurketaz jasotzen dugun ezagutzaren eta informazioaren arduraduna dena. Oro har, testu-ulermena lantzeko eta ebaluatzeko aukera anitzeko galderak eta galdera irekiak erabili ohi dira. Dudarik gabe, metodo bakoitzak abantaila zein desabantaila batzuk dakartza; esaterako, kontzeptuak modu sendoan gureganatzeko aukera anitzeko galderen eraginkortasuna zalantzagarritzat jo da (Davies, 2002; Conole eta Warburton, 2005), ezagutza hautemateko test modukoetan oso erabiliak badira ere. Galdera irekiei dagokienez, hauek erantzutea esfortzu handiagoa eta kontzeptuak argi antolatuak izatea eskatzen duenez, formakuntza garaian ezagutza modu eraginkorragoan finkatzea ahalbidetzen dutela argudiatu da (Karpicke eta Roediger, 2008). HS teknologiaren ikuspuntutik bai aukera anitzeko galderen baita galdera irekien sorkuntza automatikoa, ingelesez *Question Generation (QG)*, erronka handiko ataza da.

Behin hiru ikerketa-lerroak gainbegiratuta, euskara automatikoki sortzeko Ixa Taldean garatu diren oinarrizko baliabideak azalduko ditugu. Esaterako, EDBL datu base lexikalak (Aldezabal *et al.*, 2001) hitzak ematen dizkigu, Euskararen morfologiaren deskribapenari esker (Alegria *et al.*, 1996), hitzen sorkuntza egin dezakegu, hitzaren lema eta ezaugarri morfologikoak emanda. Euskal Wordnet (Pociello, 2008) erabilita, hitzen arteko erlazio semantikoak azter ditzakegu eta bertatik sinonimoak, hiponimoak, meronimoak e.a. lortu. Semantika eredu distribuzionalekin ere landu dugu, horretarako Mikolov *et al.* (2013) lana erabili dugu. E-Roldak (Estarrona, 2014) aditz ezberdinen informazio semantikoa eta bere argumentuen rol semantikoak eskaintzen dizkigu, beharrezkoa "*Autoa atxilotu naiz.*" bezalako esaldien sorkuntza ekiditeko. Egitura sintaktikoak aukeratzeko, corpus azterketa batean oinarritzen gara (Gonzalez-Dios *et al.*, 2015). Testuak analizatzeko Ixa Taldeak garatu duen analisi katea (Aduriz *et al.*, 2004) erabiltzen dugu.

Laburbilduz, HSak zeresan handia dauka HPko aplikazio aurreratuetan. Tekonologia hau erabiltzen duten aplikazioek, gainera, gaurkotasun nabarmena dutela azpimarratu nahi dugu.

3 Ikerketaren muina

Euskararen sorkuntza automatikoa hiru ikerketa-lerrotan aplikatu da batez ere: poesiaren sorkuntzan, testuen sinplifikazioan eta galderen sorkuntzan. Atal honetan HSa ikerketa-lerro horietan nola erabiltzen den azalduko dugu.

Edozein testuren sorkuntza burutzeko, modu edo eredu ezberdinak aurkeztu dira. Lehen aitatu bezala, Reiter eta Dale autoreek (2000) proposatutako HSrako arkitekturan oinarrituko gara bertso-sortzailearen diseinurako. Beraz, hura eraikitzeko, modulu ezberdinez osatutako sistema bat diseinatu dugu. Modulu hauetan honako funtzioak burutu behar dira: i) edukiaren zehaztapena (zer esan nahi dugu?), ii) dokumentuaren egituratzea (nola eta zein ordenatan esan nahi dugu?), iii) lexikalizazioa (zein hitz zehatzekin esango dugu?), iv) agregazioa (elkar al ditzakegu antzeko esaldiak? "*Jonek arroza jan du*" eta "*Mirenek arroza jan du*" esaldiak elkar ditzakegu, "*Jonek eta Mirenek arroza jan dute*" lortuz, v) aipamenen sorkuntza (testuan "*Santa Klara*" beharrean "*Donostiako irla*" jar genezake) eta vi) azaleko sorkuntza (aurreko ataletan sortutako egitura abstraktuak esaldi arrunt bilakatzeko prozesua).

Gai bat emanda bertsoa sortzea da gure helburua. Bertsoak sortzerakoan, gai batekin erlazionatutako puntu independenteak sortuko ditugu, eta gero hauek ordenatu egingo ditugu. Une honetan, gai gara puntu bakoitzeko beharrezko informazio minimoa sortzeko.

Puntuak sortzeko, lehenik eta behin, esan beharrekoa erabakitzeak berebiziko garrantzia du, edukiaren zehaztapenak, alegia. Idatzi beharreko hori Egitura Abstraktu Sintaktikoetan (EAS) gordeko dugu. Gure kasuan, EAS bakoitzak puntu bat errepresentatuko du. Bertso sorkuntzarako sinplifikazio bat egingo dugu, puntu bakoitzak aditz bakarra duela suposatuko dugu. Hala gertatzen da bertsoen corpusean

(Xenpelar Dokumentazio Zentroa, 2007) lerro kopuru bikoitia duten bertsoetako puntuen % 44,69tan[3]. Gaiarekin erlazionatutako hitzak lortzeko berriz, semantika distribuzionaleko teknikak erabiliko ditugu[4]. Aditzak lortzeko, izen eta aditzen arteko erlazio estatistikoak kalkulatuko ditugu (adibidez, "lan" izenarekin "deitu" edo "haserretu" bezalako hitzak proposatzen zaizkigu). Horrela, puntu bakoitzeko, izen-aditz bikote bat lortuko dugu.

Hauek lortuta, EAS bakoitzean esaldia osatu ahal izateko informazio minimoa gehituko dugu. Adibidez, "*haserretu*" aditzak bi argumentu izan ditzake, lehenengoa (arg0), kausa, ergatiboan edo soziatiboan eta bigarrena (arg1), jasotzaile/nozitzailea, absolutiboan (gizaki izan behar da arg1 elementua). Honen ondorioz, "*lan*" hitza arg0 posizioan joango da eta arg1 posiziorako hitz bat bilatu beharko dugu, adibidez, "*ni*". Horrela, {aditza, arg0, arg1} eskeman {*haserretu, lan, ni*} balioak jarri genitzake. Hau da puntu bat sortzeko behar dugun informazio minimoa.

Testuen sinplifikazioari erreparatuz, testutik testurako sorkuntza egiten da ikerketa-lerro honetan oro har. Hau da, testu bat hartuta eta sintaxia eta lexikoa moldatuz, testu berri bat sortzen da. Hala ere, sintaxiaren aldaketetan testuaren edukia eta esanahia gal ez dadin, hitz berriak gehitu egiten dira eta jada baliagarriak ez diren hitzak edo morfema funtzionalak (determinatzaileak, izenordainak, menderagailuak, kasu markak...) ezabatu egiten dira. Euskaraz, perpaus adberbialetan menderagailuak eta kasu markak ezabatzen dira eta horiek ematen duten informazioa mantentzeko adberbioak edo izen sintagmak gehitzen dira. Eragiketa honetan aditzen sorkuntza egin behar da, mendeko aditz izatetik aditz nagusi bilakatzen baitira.

Azaldu dezagun behar honen arrazoiak adibide batekin. Demagun, *ikusi dudanean* aditza aurkitzen dugula testuan. Aldiberekotasuna adierazten duen denbora perpaus bat denez, dagokion sinplifikazioko erregelan adierazten zaigu menderagailua eta kasu marka ezabatu behar ditugula aditz laguntzailetik *dudanean* eta *ordu* + inesibo kasu marka (*orduan* adberbiboa) gehitu behar dugula perpaus nagusian. *-(e)n* menderagailua eta *-(e)an* kasu markak ezabatuta, **dud* forma da geratzen dena. Forma hori ez da zuzena, *dut* izan beharko litzateke. Nola lortu hori? Alde batetik, erregela fonotaktikoak erabil genitzake *-d-* hori *-t* bihurtzeko, baina, bestetik, HSko baliabideak erabil ditzakegu aditz laguntzailea sortzeko. Beraz, hitz horren analisian oinarrituko gara, ordenagailu errepresentazio bezala, aditzaren moldaketa egiteko. Horretarako, 1. irudian ikusten den *dudanean* hitzaren analisiari INE (inesiboa) kenduko diogu eta morfologia sortzaileari (Alegria *et al.*, 1996) analisi berria emanda, *dut* aditz laguntzailea sortuko du.

1 Irudia: *dudanean* eta *dut* aditz laguntzaileen analisi morfologikoa

```
"dudanean"  "*edun"  ADL INE A1 NR_HURA NK_NIK
"dut"   "*edun"  ADL A1 NR_HURA NK_NIK
```

TSan egin dugun beste sorkuntza mota bat patroietan oinarritutakoa da. Sorkuntza mota hau arkaikoagoa den arren, datu biografikoak parentensi bidez ematen dituzten testuetan oso baliagarria da. *Biografix* (Gonzalez-Dios *et al.*, 2014) izeneko tresna independiente eleanitza da, datu biografikoak esaldietan ematen dituena. Tresna honek datu biografikoak modu irakurterrezagoan ematea eta paretesiek suposatzen duten etena ezabatzea du helburu. 1. adibidean *Biografix*ek esaldiak nola sinplfikatzen dituen ikus dezakegu.

(1) • **Jatorrizko esaldia:** Ernest Rutherford, Nelsongo lehenengo baroia, (Brightwater, Zeelanda Berria, 1871ko abuztuaren 30a - Cambridge, Ingalaterra, 1937ko urriaren 19a) fisika nuklearraren aita izan zen.
  - **Sinplifikatutako esaldiak:**
  - Ernest Rutherford, Nelsongo lehenengo baroia, fisika nuklearraren aita izan zen.
  - Ernest Rutherford 1871ko abuztuaren 30ean, Brightwateren jaio zen.
  - Brightwater Zeelanda Berrian dago.
  - Ernest Rutherford 1937ko urriaren 19an, Cambridgen hil zen.

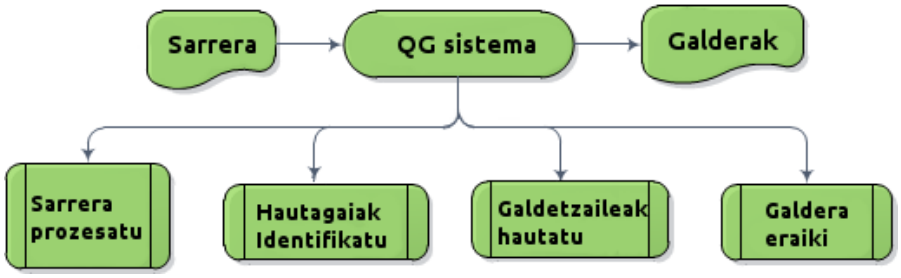
[3]Neurri handiko bertsoetako puntuen % 36,33 zen aditz bakarrekoa.
Neurri txikiko bertsoetako puntuen % 51,56 zen aditz bakarrekoa.

[4]Semantika distribuzionala metodo distribuzionalen bidez, hitzen arteko antzekotasunak neurtu eta kategorizatzea helburu duen ikerkuntza arloa da.

– Cambridge Ingalaterran dago.

Galderen sorkuntza automatikoari dagokionez (Lopez-Gazpio, 2013), gehiengoaren hizkuntzekin alderatuta lehen urrats hutsalak badira ere, euskararako teknologia hau lantzen hasiak gara ikertzaileok. Esaterako, Aldabe *et al.* autoreek 2006 urtean zenbakien inguruan emandako galdera sorkuntza automatikoaren urratsei jarraiki lau urratsetan oinarritutako galdera sorkuntza sistema garatu zuten. Arkitektura honen oinarrizko urratsak ondorengoak dira: i) sarrera fitxategia prozesatzea, ii) galderaren emaitza izango den hautagaia aukeratzea, iii) galdetzaileak (*n*or, zer, non, noiz...) identifikatzea eta iv) galdera eraikitzea. Aipatutako urratsak 2. irudian ikus daitezke grafikoki azalduta.

2 Irudia: Galderak sortzeko prozedura.



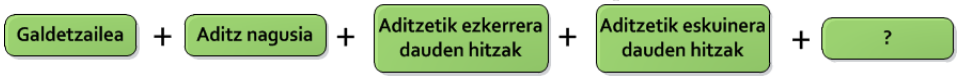
Lehen fasean sarrera testuaren inguruko ezaugarri linguistikoak biltzen dira: izenak, izen lagunak, aditzak, aditzondoak, etab. Oro har, prozesu hau hainbat azpi urratsetan egikaritzen den ataza da, eta azpi urrats hauetan HPko tresnak erabiliz sarrera corpusa (testu multzoa) analizatu eta informazio linguistikoa biltzen da. Behin tresna hauetatik abiatuta, informazio guztia jasotzean, sistema hurrengo atazekin hasteko gai da.

Galderen emaitzak izango diren hautagaiak identifikatzeko irizpidea lehiaketa batean oinarritzen da, lehiaketaren hautagaiak esaldiko izen sintagma guztiak izanik. Hala, esaldi bakoitzeko izen sintagma guztiak sekuentzialki prozesatzen dira hautagairik garrantzitsuenak bilatzeko asmoz. Hautagaien interesa neurtzeko pisutze sistema bat erabiltzen da, eta pisua hautagaiarekin erlazionatuta biltzea lortu den informazioaren araberakoa da. Sistema honen kasuan hainbat informazio iturri erabiltzen dira, alde batetik, izen sintagmen informazio linguistikoa, eta, bestetik, HPko tresnak identifikatzeko gai diren informazio semantiko gehigarria, hala nola: hautagaien biziduntasuna, toki edo leku informazioa, hautagaiak esaldian duen rol semantikoa eta abar.

Hautagaiak identifikatu ostean, prozesuaren hirugarren atalari ekiten dio sistemak. Urrats honen helburua galdetzailearen edo galdera motaren identifikazio zuzena da. Galdetzaileen aukeraketa hautatu diren izen sintagmen informazio linguistikoaren arabera egiten da. Honela, ezaugarri linguistikoaren arabera hizkuntzalariek ezarritako galdetzaileak erabiltzen dira, dagozkien izen sintagmentzako probabilitate handiz zuzenak izateko estimatu direnak.

Azkenik, aplikazioak galderaren sorkuntzari ekiten dio. Fase hau egikaritzeko, algoritmo bat erabiltzen da, esaldiko sintagmak berrantolatzen eta esaldiari galderazko zentzua ematen diona; ahalik eta galdera naturalena lortzeko helburuarekin. Eraldaketa patroi hau 3. irudian ikus daiteke.

3 Irudia: Galdera eraikitzeko patroia.



Prozesu honen egikaritzapena 2. adibidean ikus daiteke. Jatorrizko esaldia emanda, galdera sortzaileak ondorengo galdera sortuko du.

(2) • **Jatorrizko esaldia:** Toxinekiko erresistentzia beste landare batzuetara zabal daiteke.
– **Sortutako galdera:** Nora zabal daiteke toxinekiko erresistentzia?

Egitura konplexuagoa duten esaldietan galdera sortzaileak zailtasunak ditu galdera zentzudunak osatzeko. Testu sinplifikatzaileak aurreprozesu gisa erabilita, kalitate handiagoko galderak lortzen dira. 3.

adibidean, *Biografix* aplikatu aurretik eta ondoren lortzen diren emaitzak ikus ditzakegu.

(3)
- **Jatorrizko esaldia:** Eduardo Hughes Galeano (Montevideo, 1940ko irailaren 3a -) Uruguaiko kazetari eta idazlea da.
  - **Sortutako galdera:** Nor da Eduardo Hughes Galeano Montevideo 1940ko irailaren 3a?
- **Sinplifikatutakoa:** Eduardo Hughes Galeano Uruguaiko kazetari eta idazlea da. Eduardo Galeano 1940ko irailaren 3an Montevideon jaio zen.
  - **Sortutako galdera:** Nor jaio zen 1940ko irailaren 3an Montevideon?
  - **Sortutako galdera:** Non jaio zen Eduardo Galeano 1940ko irailaren 3an?

Atal honetan ikusi ahal izan dugunez, egun HS euskaraz hiru ikerketa-lerro ezberdinetan aplikatzen da. Hala ere, ikerketa-lerro honetan sustrai berriak errotuko direlakoan gaude.

4 Ondorioak

Artikulu honetan euskararen sorkuntza automatikoa nola gauzatzen ari den azaldu dugu. Horretarako, atzerriko hizkuntzetan egin dituzten lanak aipatu ondoren, HS euskaraz hiru ikerketa lerroetan nola aplikatzen den azaldu dugu.

Poesiaren sorkuntzan, gai bati testu batekin nola erantzungo diogun aurkeztu dugu, hainbat moduluz osatutako arkitekturaren hastapenak erakutsita. Modulu arteko komunikaziorako egitura abstraktu sintaktikoak erabiltzen ditugu eta hauekin puntuak sortuko ditugu etorkizun hurbilean.

Testuen sinplifikazioan, aditz nagusiak nola sor daitezkeen azaldu dugu batetik, eta, bestetik, *Biografix* tresna aurkeztu dugu. Aipatu ditugun lan horiek testuak automatikoki sinplifikatzeko idatzi ditugun erregela eta proposamenetako batzuk besterik ez dira. Egitura parentetikoekin egiten dugun bezala, perpaus adberbialetatik, erlatibozko perpausetatik eta aposizioetatik ere esaldi berriak sortzen ditugu, egitura linguistiko konplexuak aztertu ondoren, kasuz kasuko berridazketak eta sorkuntzak egiten baititugu.

Galderen sorkuntzan, testu fitxategi batetik abiatuta euskaraz galderak sortzeko gai den sistema baten hastapena deskribatu dugu; hainbat hizkuntza baliabide erabiliz galderak egiten saiatzen dena. Ebaluazioko emaitzetan ikusi dugu % 40 inguru zuzena dela (Aldabe *et al.*, 2013). Gainera, ehuneko hori igo edo jeisten da, erabili den heuristikoaren arabera. Oro har, galderen sorkuntza ikerketa-lerro zabala da, eta, dagoeneko, azaleko sintaxiaz gain hitzen arteko dependentzietan oinarrituta galderak sortzeko gai diren sistema konplexuagoak ere garatu dira, esaterako, Madrazo Azpiazu autoreak 2013 urtean garatutako galdera sorkuntza sistema.

5 Etorkizunerako planteatzen den norabidea

Etorkizunerako hainbat erronka eta soluzio posible aurreikusten ditugu HS euskaraz egiten ari garenak. HPko gainontzeko aplikazio aurreratuetan sor daitezkeen arazoei aurre egiteaz gain, aztertu ditugun hiru ikerketa-lerroetan badago zer egin.

Poesia-sorkuntzan, etorkizun hurbilean egin beharreko lana sistemaren inplementazioa bukatzea da. Oraingoz, gai bat emanda puntuetarako informazio minimoa sortzeko gai gara. Informazio horretatik abiatuta esaldi sintaktikoki eta metrikoki zuzenak sortzea da helburua. Helburu honen bidean puntuen metrika lantzeko Agirrezabal *et al.* autoreen 2012 laneko tresnak erabiliko ditugu, puntuen ordenaziorako Lapata (2003) ikerlanean aurkeztutako teknikak eta azkenik, azaleko sorkuntzarako Agirrezabal *et al.* autoreen (2013) laneko proposamena.

Testuen sinplifikazioan dugun erronka nagusia sistemaren ebaluazioa da. Izan ere, hizkuntza sortu duten sistemen ebaluazioa ez dago argi. Batetik, gizakiek ebalua ditzakete, baina ebaluazio hauek garestiak izaten dira (testak egiteko azpiegitura egokiak behar dira, oso ondo pentsatu behar da zeintzuk diren egiten diren galderak edo probak e.a.) eta beste faktore batzuek eragina izan dezakete (nekea, bakoitzaren jakintza...). Bestetik, eskuz sortutako sinplifikazioen kontra ere ebalua daiteke (*Gold Standard* edo urrezko patroi balira bezala), baina sinplifikazio horiek ere sortu behar dira. Ondoren, itzulpen automatikoak edo laburpen automatikoak ebaluatzeko erabiltzen diren metrikak aplikatzen dira, baina ez dago oso garbi oraindik metrika horiek testuen sinplifikaziorako baliagarriak ote diren.

Galderen sorkuntzan argi dago etorkizuneko lana heuristikoen hobekuntzari lotuta bideratu behar dela. Hau da, helburua galderen sorkuntza hobetuko duten HPko tresna berrien analisi eta integrazioan arreta jartzea. Jakina da zenbat eta linguistikoki aberatsagoa den analisia izan, orduan eta galdera hobeak sortuko direla.

Ikus dezakegunez, erronkarik ez da falta euskararen sorkuntza automatikoan. Eta hau hasiera besterik ez da. Nora irits gintezke? Egingo al dute makinek euskaraz?

Erreferentziak

Aduriz, Itziar, María Jesús Aranzabe, Jose Mari Arriola, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, eta Larraitz Uria. 2004. A cascaded syntactic analyser for Basque. *Computational Linguistics and Intelligent Text Processing* 124–134.

Agirrezabal, Manex, Iñaki Alegria, Bertol Arrieta, eta Mans Hulden. 2012. Finite-state Technology in a Verse-making Tool. In *10th International Workshop on Finite State Methods and Natural Language Processing*, 35–39.

——, Bertol Arrieta, Aitzol Astigarraga, eta Mans Hulden. 2013. POS-tag Based Poetry Generation with WordNet. In *Proceedings of the 14th European Workshop on Natural Language Generation*, 162–166.

Aldabe, Itziar, Maddalen Lopez de Lacalle, Montse Maritxalar, Edurne Martinez, eta Larraitz Uria. 2006. Arikiturri: an Automatic Question Generator Based on Corpora and NLP Techniques. In *Intelligent Tutoring Systems*, 584–594. Springer.

——, Itziar Gonzalez-Dios, Iñigo Lopez-Gazpio, Ion Madrazo, eta Montserrat Maritxalar Anglada. 2013. Two Approaches to Generate Questions in Basque. *Procesamiento del Lenguaje Natural* 51.101–108.

Aldezabal, Izaskun, Olatz Ansa, Bertol Arrieta, Xabier Artola, Aitzol Ezeiza, G. Hernández, eta Mikel Lersundi. 2001. EDBL: a General Lexical Basis for the Automatic Processing of Basque. In *Proceedings of the IRCS Workshop on linguistic databases*.

Alegria, Iñaki, Xabier Artola, Kepa Sarasola, eta Miriam Urkia. 1996. Automatic Morphological Analysis of Basque. *Literary and Linguistic Computing* 11.193–203.

Aranzabe, María Jesús, Arantza Díaz de Ilarraza, eta Itziar Gonzalez-Dios. 2012. First Approach to Automatic Text Simplification in Basque. In *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, ed. by Luz Rello eta Horacio Saggion, 1–8.

Barzilay, Regina, Noemie Elhadad, eta Kathleen R. McKeown. 2001. Sentence Ordering in Multidocument Summarization. In *Proceedings of the first international conference on Human language technology research*, 1–7. Association for Computational Linguistics.

Besold, Tarek Richard, Marco Schorlemmer, eta Alan Smaill. 2014. *Computational Creativity Research: Towards Creative Machines*. Springer.

Coch, Jose. 1998. Multimeteo: Multilingual Production of Weather Forecasts. *ELRA Newsletter* 3.

Conole, Gráinne, eta Bill Warburton. 2005. A Review of Computer-assisted Assessment. *Research in Learning Technology* 13.17–31.

Davies, Phil. 2002. Theres no Confidence in Multiple-Choice Testing,.... In *Proceedings of 6th CAA Conference*, 119–130. Loughborough University.

Estarrona, Ainara, 2014. *EPEC corpusa predikatu-mailan etiketatzeko oinarriak: EPEC-RolSem, BVI eta e-ROLda*. Euskal Herriko Unibertsitatea (UPV/EHU) tesia.

Gervás, Pablo. Composing Narrative Discourse for Stories of Many Characters: A Case Study over a Chess Game. *Literary and Linguistic Computing* .

Goldberg, Eli, Norbert Driedger, eta Richard I. Kittredge. 1994. Using Natural-language Processing to Produce Weather Forecasts. *IEEE Expert* 9.45–53.

Gonzalez-Dios, Itziar. 2014. Euskarazko testuak errazten: euskal testuen sinplifikazio automatikoa. In *Euskal Hizkuntzalaritzaren egungo zenbait ikerlerro. Hizkuntzalari euskaldunen I. topaketa*, ed. by Itziar Aduriz eta Ruben Urizar, 135–149. Udako Euskal Unibertsitatea.

——, María Jesús Aranzabe, eta Arantza Díaz de Ilarraza. 2013. Testuen sinplifikazio automatikoa: arloaren egungo egoera. *Linguamática* 5.43–63.

——, ——, eta ——. 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)*, 11–20, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

——, ——, eta ——. 2015. Perpaus adberbialen agerpena, maiztasuna eta kokapena EPEC-DEP corpusean. Technical report, University of the Basque Country (UPV/EHU) UPV/EHU/LSI/TR 02-2015.

Jurafsky, Dan, eta James H. Martin. 2000. *Speech & Language Processing*. Pearson Education India.

Karpicke, Jeffrey D., eta Henry L. Roediger. 2008. The critical Importance of Retrieval for Learning. *Science* 319.966–968.

Lapata, Mirella. 2003. Probabilistic Text Structuring: Experiments with Sentence Ordering. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 545–552. Association for Computational Linguistics.

Lopez-Gazpio, Iñigo, 2013. Seneko: galderak automatikoki sortuz testuak lantzeko aukera ematen duen aplikazioa.

Madrazo Azpiazu, Jon. 2013. Hizkuntzaren prozesamendurako teknikak irakaskuntza arloan: galdera sortzaile automatikoa.

Mikolov, Tomas, Kai Chen, Greg Corrado, eta Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Oliveira, Hugo Gonçalo. 2012. Poetryme: a versatile platform for poetry generation. In *Proceedings of the ECAI 2012 Workshop on Computational Creativity, Concept Invention, and General Intelligence*.

——, eta Amílcar Cardoso. 2015. Poetry Generation with PoeTryMe. In *Computational Creativity Research: Towards Creative Machines*, 243–266. Springer.

——, Raquel Hervás, Alberto Díaz, eta Pablo Gervás. 2014. Adapting a Generic Platform for Poetry Generation to Produce Spanish Poems. In *5th International Conference on Computational Creativity, ICCC*, 63–71.

Pociello, Eli, 2008. *Euskararen ezagutza-base lexikala: Euskal WordNet*. Euskal Herriko Unibertsitatea (UPV/EHU) tesia.

Reiter, Ehud, eta Robert Dale. 2000. *Building natural language generation systems*. Cambrigde University Press.

Shardlow, Matthew. 2014. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing* 58–70.

Siddharthan, Advaith. 2014. A Survey of Research on Text Simplification. *International Journal of Applied Linguistics* 259–98.

Stent, Amanda, eta Srinivas Bangalore (eds.) 2014. *Natural Language Generation in Interactive Systems*. Cambridge University Press.

Toivanen, Jukka M., Oskar Gross, eta Hannu Toivonen. 2014. The Officer Is Taller Than You, Who Race Yourself! Using Document Specific Word Associations in Poetry Generation. In *Proceedings of 5th International Conference on Computational Creativity, ICCC*.

Turing, Alan M. 1950. Computing Machinery and Intelligence. *Mind* 433–460.

Turner, Ross, Somayajulu Sripada, Ehud Reiter, eta Ian P. Davy. 2006. Generating Spatio-temporal Descriptions in Pollen Forecasts. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, 163–166. Association for Computational Linguistics.

Xenpelar Dokumentazio Zentroa, 2007. Bertsolaritzaren datu-basea. http://bdb.bertsozale.eus.

Yan, Rui, Han Jiang, Mirella Lapata, Shou-De Lin, Xueqiang Lv, eta Xiaoming Li. 2013. I, poet: Automatic Chinese Poetry Composition through a Generative Summarization Framework under Constrained Optimization. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2197–2203. AAAI Press.

Zhang, Xingxing, eta Mirella Lapata. 2014. Chinese Poetry Generation with Recurrent Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 670–680.

Eskerrak eta oharrak

Manex Agirrezabalen eta Itziar Gonzalez-Diosen lanak garatzen ari diren doktoretza-tesietan oinarrituta daude. Tesia egiteko Manex Agirrezabal Zabalduz bekaren onuraduna da eta bere zuzendariak Bertol Arrieta eta Mans Hulden dira. Itziar Gonzalez-Diosek Eusko Jaurlaritzako Predok beka dauka eta bere zuzendariak Arantza Díaz de Ilarraza eta María Jesús Aranzabe dira. Iñigo Lopez-Gazpioren lana Informatikan Ingeniaritza lortzeko Karrera Bukaerako Proiektuan oinarrituta dago eta bere zuzendaria Montserrat Maritxalar izan zen. Egun, Iñigo doktoretza-tesia egiten ari da, FPU beka baten onuraduna da eta bere zuzendariak Eneko Agirre eta Montserrat Maritxalar dira.

Ikerketa hau Eusko Jaurlaritzak IXA taldea, A motako ikertalde finkatua (IT344-10), emandako diru laguntzari esker egin da. Ez genuke ahaztu nahi Euskal Herriko Bertsozale Elkartea eta bertako Xenpelar Dokumentazio Zentroa. Beraien laguntzarik gabe ez litzateke posible izango bertsoen analisiaren eta sorkuntzaren inguruan egindako lana.