Hizkuntzaren prozesamendurako teknikak irakaskuntza arloan: galdera sortzaile automatikoa

Gradu Amaierako Proiektua

2013.eko uztailaren 14

Ion Madrazo Azpiazu

Zuzendaria:
Montse Maritxalar Anglada

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Esker onak

Eskerrik asko **Montse** proiektuko zuzendari bikaina izateagatik. Momentu zailetan beti laguntzeko prest sentitu zaitut. Eskerrik asko ere ikerketaren mundu hau ezagutzera emateagatik.

Eskerrik asko **Itziar** emandako laguntza guztiagatik, proposatzen zenituen ideia zein azaldu zenizkidan kontzeptu estatistiko guztiengatik.

Eskerrik asko **Iñigo** irakurketa taldeko kide bikaina izateagatik. XML tresnekin emandako laguntza ere primerakoa izan zen.

Eskerrik asko **Itziar Gonzalez** galderen ebaluazioan emandako laguntzagatik. Momentu zailak izan ziren arren aurrera ateratzea lortu genuen.

Eskerrik asko **Zezilio** burutu duzun primerako ebaluazioaz gain, txikitatik euskarako irakasgai hartan hizkuntzalaritzan eman zenidan formazioagatik. *Zeinek esango zin nei zuk erakutzitteko kontu haxe danak gradu bukaerako proiektun ibiliko nittunik!*

Eskerrik asko **Aitor** Wikipedian prozesatzeko eman didazun laguntzagatik, zu gabe ezingo nuen hau egin.

Eskerrik asko **Bertol** perpaus banaketan eman zenidan laguntzagatik.

Eskerrik asko **Kepa** proiektua androiderako integratzea proposatu izanagatik, zuzentzaile ortografikoengatik eta nire bizitzako lehen programa idazten erakusteagatik. Mundu berri bat ireki zenidan honekin.

Eskerrik asko **Ixa ikerketa taldeko** kide guztioi. Egunero burutzen duzuen lana guztiz beharrezkoa da hainbeste lan ematen digun hizkuntza hau garai berrietara egokitzeko.
Eskerrik asko **unibertsitateko lagunoi** elkarrekin genituen duda guztiak

argitzea lortu dugulako. Zuek gabe ziur aski ez nuen proiektua matrikulatzerik ere lortu izango ;)

Eskerrik asko **Mikel** ikasketetan zehar izan ditudan blokeo guztietatik ateratzen lagundu nauzulako. Zugatik izan ez balitz ziur aski ez nintzen informatikako graduan hasiko.

Eskerrik asko nire lagunmin **Ioni** proiektuan zehar aguantatu nauzulako eta txikitatik nire alboan egon zarelako. Proiektuan zehar egon ez garen aldi guztiak berreskuratuko ditugu uda honetan.

Azkenik, eskerrik asko nire **familiari**. Askotan zuek konturatu ez arren zuen babesak aurrera jarraitzen lagundu didalako.

Gaien Aurkibidea

Irudien Zerrenda

Taulen Zerrenda

Laburpena

Jarraian aukezten den dokumentuan irakaskuntza arloan laguntzeko sistema baten azterketa eta garapena azaltzen dira. Zehatzago hitz eginez, bi moduluz osaturiko galdera sortzaile automatiko bat aurkezten da. Batetik, testuko esaldi esanguratsuak detektatzeko tresna baten garapenaren berri ematen da. Bestetik, aukeratutako esaldien gainean galderak eraikitzeko gai den sistema bat azaltzen da. Modulu bien zein eszenatokiko ebaluazioak garatzeko burutu den corpus azterketa bat ere aurkezten da.

1 Kapitulua

Proiektuaren Helburu Dokumentatua

Gaien Aurkibidea

1.1 Proiektuaren deskribapena eta helburuak

Proiektu honen helburua hizkuntzaren prozesamendurako tresnek irakaskuntza arloan izan dezaketen erabilgarritasuna aztertzea da. Konkretuki, irakaskuntza materialen sorkuntzan laguntza handia eskaini dezake gaur egun hizkuntzaren prozesamenduak.

Ariketak automatikoki prestatzeko sistemak, testu idatzien kalitatea hobetzen laguntzeko sistemak, laburpengintza sistemak... denetarik sortu da azken aldian. Baina, beste hizkuntzetarako buruturiko lana handia den arren euskararako buruturiko lana oso murritza da. Honek bultzatuta, proiektu honetan euskararako ariketak automatikoki sortzeko sistemetan azterketa bat burutuko da, galdera ariketak automatikoki sortzeko zehazki. Galdera hauen helburua ikasleek testuen ulermena lantzea izango da, horretarako testuko alderdi esanguratsuenei buruz galdetzen saiatuko direlarik. Ataza hau burutzeko bi ikerketa lerrotan sakonduko da:

Alde batetik, testuko zati esanguratsuenak bilatzen lagunduko duen modulu bat inplementatuko da. Modulu honen helburua testuan agertzen diren termino garrantzitsuenak markatzea eta pisatzea izango da. Termino esanguratsuak bilatuz testuan garrantzitsu diren kontzeptuak zein izan daitezkeen jakin dezakegu. Hau abiapuntu egokia izan daiteke testu baten ulermena lantzen hasteko. Termino hauek markatzeko metodoa probabilitatean oinarriturikoa izango da, hizkuntzaren prozesamenduko alderdi estatistikoa jorratuz.

Beste alde batetik, testuan bilaturiko termino garrantzitsuenen inguruan galderak automatikoki eraikiko dituen modulu bat inplementatuko da. Honen helburua esaldi bat eman eta ahalik eta galdera zentzuzko eta zuzenenak sortzea izango da. Honetarako erregeletan oinarrituriko modulu bat inplementatuko da, hizkuntzaren prozesamenduko alderdi linguistikoa jorratuz.

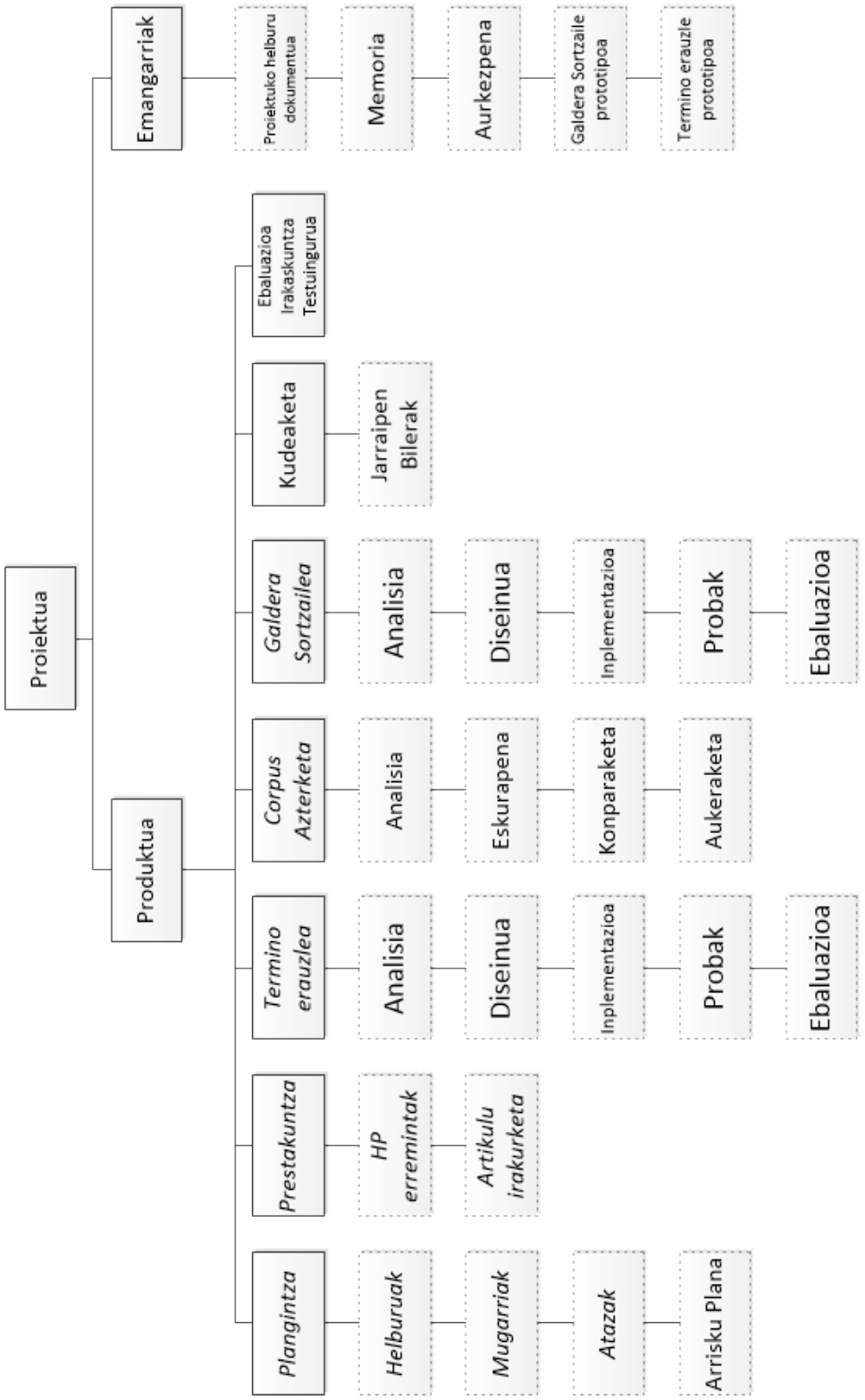
Bi modulu hauekin probak burutu ahal izateko euskararako existitzen diren hainbat corpusen azterketa burutuko da. Proiekturako interesgarriak izan daitezkeen corpusak eskuratu eta beharrezko bada corpus berriak biltzeko helburua izango du azterketa honek.

Amaitzeko, modulu bakoitzean modu independentean burutuko diren ebaluazioez gain ebaluazio orokor bat ere burutuko da. Bertan bi moduluak elkarrekin lotu eta eszenatoki erreal batean sorturiko aplikazioak izan dezakeen erabilgarritasuna aztertuko da.

1.2 Proiektuaren plangintza

1.2.1 Lanaren deskonposaketa egitura (LDE)

Jarraian proiektuaren lanaren deskonposaketa egitura (LDE) aurkezten da. Bertan proiektuan zehar burutu beharreko lan karga ataza eta azpiataza garrantzitsuenen arabera erakusten da.



Irudia 1.1: Lanaren deskonposaketa egitura

1.2.2 Atazen definizioa

Jarraian proiektuaren lan pakete edo ataza ezberdinak aurkeztu eta deskribatu egiten dira:

- **Plangintza**
  Ataza honetan proiektua garatzeko beharrezko izango den planifikazio bat garatuko da. Bertan proiektuaren helburuak finkatzeaz gain berau lan paketetan banatu eta denboraren arabera estimatuko da. Proiektuaren bideragarritasun eta arrisku plana ere bertan garatuak izango dira.

- **Prestakuntza**
  Ataza honetan proiektua garatzeko beharrezko izango den prestakuntza barneratuko da. Artikuluak irakurri, aurrekariak aztertu, zein beharrezkoak izango diren tresna informatikoetan trebezia lortzeko balioko du ataza honek.

- **Galderak sortzeko modulua**
  Ataza honetan galdera motaren aukeraketa eta eraikuntzaz arduratuko den modulua sortzeaz gain, existitzen diren beste sistemekiko konparaketa bat gauzatuko da. Ataza hau edozein proiekturen bizitza zikloa osatzen duten azpiatazez osatua dago:

  - Analisia
  - Diseinua
  - Inplementazioa
  - Probak
  - Ebaluazioa

- **Corpus azterketa**
  Ataza honen helburua euskararako existitzen diren corpusen inguruko azterketa bat burutzea du helburu.

  - Analisia
    Ataza honetan aukeratutako domeinuen inguruan existitzen diren corpusak aurkituko dira.
  - Eskurapena
    Ataza honetan aurkitu diren corpusak eskuratzeko eta proiekturako interesgarria izan daitekeen formatuan uzteko lana egingo da.

  - Konparaketa
    Behin corpusak eskuratuta beraien arteko konparaketa bat gauzatuko da.
  - Aukeraketa
    Konparaketatik ateratako ondorioen arabera corpus baliagarrienak aukeratuko dira.

- **Termino erauzlea**
  Ataza honetan eduki aukeraketaz arduratzen den modulua sortzeaz gain, existitzen diren beste sistemekiko konparaketa bat gauzatuko da. Ataza hau edozein proiekturen bizitza zikloa osatzen duten azpiatazez osatua dago:
  - Analisia
  - Diseinua
  - Inplementazioa
  - Probak
  - Ebaluazioa

- **Ebaluazioa irakaskuntza testuinguruan**
  Ataza honetan, behin galdera sortzailearen modulu guztiak bukatuta eta lotuta daudela, irakaskuntza ingurune batean ebaluazio bat gauzatuko da. Ebaluazio honek sistema osoaren baliagarritasuna frogatzeko balioko du.

- **Jarraipen eta kontrola**
  Ataza hau proiektua bizirik dirauen bitartean aktibo egongo da. Proiektua denboraz zein helburuz egoki dabilela kontrolatzeko balioko du. Honetarako zenbait jarraipen bilera burutuko dira.

- **Dokumentazioa**
  Ataza honen helburua proiektuaren dokumentazioa idaztea da.

- **Aurkezpenaren prestaketa**
  Azken ataza izango da hau, proiektu osoa bukatu eta gero aurkezpen egoki bat prestatzeko lana burutuko da bertan, proiektuaren defentsa modu egokia burutu ahal izateko.

1.2.3 Emangarriak

Bost emangarri identifikatu dira proiektuan. Jarraian azaltzen dira:

Memoria

Proiektuaren zeresan guztiak biltzen dituen dokumentua izango da hau. Bertan proiektuaren deskribapena, plangintza, berorren fase ezberdinen inguruko azalpenak eta bukaerako ondorioak azalduko dira. Hau osatzeko datuak proiektua aurrera doala biltzen joango dira.

Proiektuaren helburu dokumentua

Memoriaren zati bat izango da, proiektuaren hasieran garatua izango dena, bertan proiektuaren irismen zein planifikazio kontuak azalduko dira, honen bideragarritasuna ere aztertuko delarik.

Aurkezpena

Proiektuaren bukaeran berau defendatu beharko da. Defentsa hau burutzeko proiektuaren aurkezpen bat prestatu beharko da. Bertan proiektuaren ezaugarri zein arazo nagusiak azaldu beharko dira, modu labur eta ordenatu batean.

Termino erauzlearen prototipoa

Termino erauzlearen prototipoa, gauzatuko den ikerketaren ondorioz sortuko den lehen ekarpena izango da.

Galdera sortzailearen prototipoa

Galdera sortzailearen prototipoa, gauzatuko den ikerketaren ondorioz sortuko den bigarren ekarpena izango da.

1.2.4 Mugarriak

Proiektua aurrera ongi eramatea baldintzatzen duten zenbait mugarri identifikatu dira, jarraian aurkezten dira:

1.2.4.1 Barne Mugarriak

- **2013/01/21 – Galdera sortzailearen modulua bukatzeko epea**
  Egun honetarako galdera sortzailearen modulua inplementatu eta probatua egon biharko da.
- **2013/02/21 – Corpusen azterketaren bukaera**
  Egun honetarako corpusak eskura eta aukeratuta egon beharko dira.

- **2013/05/21 – Termino erauzlea bukatzeko epea**
  Egun honetarako terminoak erauzteko moduluak bukatua egon beharko du.

- **2013/05/30 – Dokumentazioarekin hasteko eguna**
  Egun honetarako proiektuko ataza nagusiak bukatuak egon beharko dira eta dokumentazio faseari ekingo zaio.

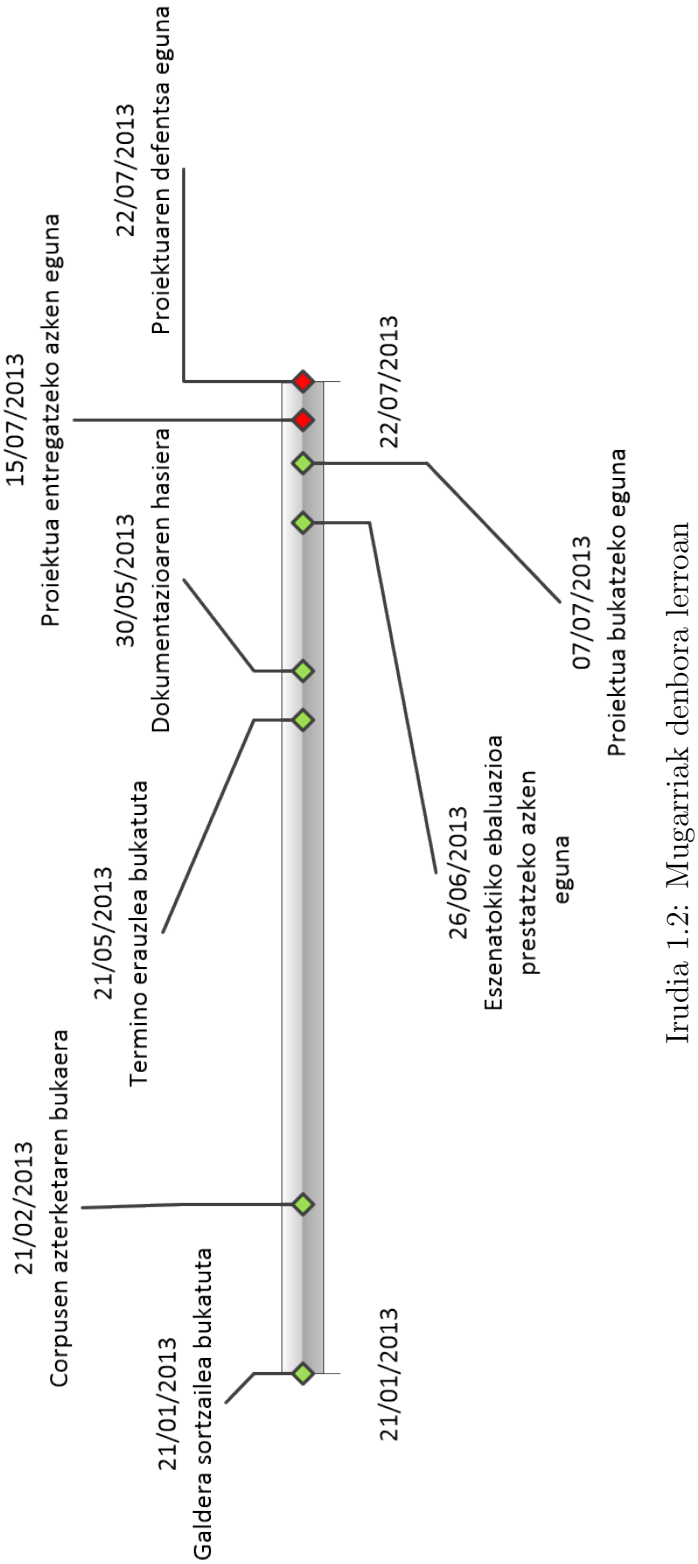
- **2013/06/26 – Eszenatokiko ebaluazioa prestatzeko azken eguna**
  Egun honetarako eszenatokiko ebaluazioak bukatuta egon beharko du.

- **2013/07/07 – Proiektua bukatzeko azken eguna**
  Egun honetarako proiektuko ataza guztiak bukatuak egon beharko dute aurkezpena prestatzearen ataza izan ezik.

1.2.4.2 Kanpo Mugarriak

- **2013/07/15 – Proiektua entregatzeko azken eguna**
  Egun honetarako proiektuko ataza guztiak bukatuak egon beharko dute aurkezpena prestatzearen ataza izan ezik.

- **2013/07/22 edo 2013/07/24 – Proiektua defenditzeko egunak**
  Egun honetarako aurkezpenak prestatua egon beharko du.



Irudia 1.2: Mugarriak denbora lerroan

1.2.5 Kronograma

Atal honetan, 1.2.2 atalean aipatutako ataza eta azpiatazek egutegian duten kokapenaren adierazpen grafikoa aurkezten da *Gantt* diagrama baten bitartez. Bertan ataza bakoitzaren hasiera eta amaiera datak finkatzen dira. Denbora estimazioen datu zehatzagoak ikustea nahi izanez gero jo I. eranskinera.

1.3 irudian ikus daiteken moduan prestakuntza atalak proiektu ia osoan zehar dirau eta beste zenbait atazarekin batera burutua izango da. Atentzioa dei dezake galdera sorkuntzaren ebaluazioaren kokapenak. Hau data horietan kokatzeak beste ebaluazioekin batera burutzea du helburu. Horrela behin garapenarekin bukatu ostean ebaluazioan guztiz zentratzeko aukera egongo da, bai ikasle zein ebaluatzaileen partetik.

Id.	Ataza	Azpiataza	Hasiera	Bukaera
1	Plangintza		07/09/2012	07/11/2012
2	Prestakuntza		17/09/2012	21/03/2013
3	Galdera Sortzailea	Analisia	07/11/2012	21/11/2012
4	Galdera Sortzailea	Diseinua	21/11/2012	30/11/2012
5	Galdera Sortzailea	Inplementazioa	03/12/2012	01/01/2013
6	Galdera Sortzailea	Probak	01/01/2013	21/01/2013
7	Galdera Sortzailea	Ebaluazioa	15/04/2013	01/05/2013
8	Corpus azterketa	Analisia	21/01/2013	01/02/2013
9	Corpus azterketa	Eskurapena	01/02/2013	15/02/2013
10	Corpus azterketa	Konparaketa	15/02/2013	19/02/2013
11	Corpus azterketa	Aukeraketa	19/02/2013	21/02/2013
12	Termino Erauzlea	Analisia	21/02/2013	07/03/2013
13	Termino Erauzlea	Diseinua	07/03/2013	15/03/2013
14	Termino Erauzlea	Inplementazioa	15/03/2013	15/04/2013
15	Termino Erauzlea	Probak	01/05/2013	07/05/2013
16	Termino Erauzlea	Ebaluazioa	07/05/2013	21/05/2013
17	Ebaluazioa irakaskuntza...		21/05/2013	07/06/2013
18	Kudeaketa		07/09/2012	22/07/2013
19	Dokumentazioa		30/05/2013	28/06/2013
20	Auzkezpen Prestaketa		15/07/2013	22/07/2013

Irudia 1.3: Gantt diagrama

1.3 Lan Metodologia

Proiektua ongi garatu ahal izateko hasieratik baldintza batzuk zehaztea beharrezkotzat jotzen da. Baldintza hauek ez dira behin betikoak, izan ere, proiektua aurrera doan heinean, eraldatuz joango dira.

1.3.0.1 Bilerak

Astero jarraipen bilerak egingo dira, egun eta ordu konkretu batzuk zehaztuz. Bilera hauek ordubete ingurukoak izatea estimatzen da, baina beti ere bertan jorratuko diren atazen araberako luzera izango dute. Hitzorduak ere aldaketak jasan ahal izango ditu ikasle edo tutorearen ezuste edo lan pilaketaren ondorioz. Aldaketa hauek e-mailez ala pertsonalki komunikatu ahal izango dira. Ezusteak albo batera, ikasle eta tutorearen partetik astero bilera bat gutxienez burutzeko esfortzua burutuko da.

1.3.0.2 Planifikatutako ordutegiak

Planifikazioaren ondorioz sortu diren epe eta ordutegiak ahal den heinean beteko dira. Kontuan izan behar da proiektuarekin batera klasea jasotzen egongo dela ikaslea, beraz klase hauek aurrera ateratzeari lehentasuna emango zaio. Hala ere epeak ahalik eta modu zorrotzenean betetzeko esfortzua burutuko da.

1.3.0.3 Prestakuntza

Prestakuntza bi modutan burutuko da. Batetik ikasleak bere partetik prestakuntza bat burutuko du, tutoreak gomendaturiko artikulu eta liburuetan oinarrituta. Bestetik bi astean behin irakurketa talde bat antolatuko da, tutorea, ikertzaile bat eta bi ikasleren artean. Aldi bakoitzean artikulu bat proposatuko da irakurtzeko eta kide guztiek irakurriko dute. Aste bakoitzean kide batek artikuluan irakurritakoa azalduko du, guztien artean sorturiko dudak argituko direlarik.

Hasierako fasean denbora gehiago dedikatuko zaio prestakuntzari, baina denbora aurrera doan heinean prestakuntzari denbora gutxiago dedikatuko zaio, irakurketa bilerak maiztasun txikiagoarekin burutuko direlarik.

1.3.0.4 Inplementazioa

Inplementazioan zehar probak burutuko dira ahalik eta sarrien. Proba fasea bukaerako ez uzten saiatuko da, horrela erroreak modu aurretiago batean zuzenduko direlakoan, eta zuzentze lana errazagoa izango delakoan.

1.4 Bideragarritasuna

Proiektua aurrera eramateko beharrezkoak izango diren baldintzak aztertu dira, ondoren zehazten dira garrantzitsuenak:

1.4.0.5 Baliabideen kostua

Proiektuan zehar beharrezkoak izango diren baliabideak doakoak direla eta proiektuak iraungo duen bitartean horrela izaten jarraituko dutela bermatu da.

1.4.0.6 Baliabideen funtzionamendu bermea

Erabiliko diren baliabideak proiektua egiteko momentuan prest eta atzigarri egongo direla bermatu da. Hala nola, oraindik garapen prozesuan daudenak proiekturako beharrezkoa diren funtzionalitateak prest dituztela bermatu da eta urrunetik atzituko diren baliabideak beti aktibo eta erabilgarri egongo direla bermatu da.

1.4.0.7 Denbora

Denbora aldetik proiektua aurrera eramateko nahikoa izango dela bermatu da plangintzaren bitartez. Proiektuan zehar garatuko diren programak garatzeko denbora izango dela bermatu da.

1.4.0.8 Komunikazioa

Ikasle eta tutorearen artean komunikazio arazoak ekiditeko lan metodologia egoki bat finkatu da 1.3 atalean. Honek desadostasunak ekidin eta proiektua aurrera ongi eta garaiz aterako dela bermatuko du.

1.5 Arrisku analisia

Proiektu batean zehar arrisku ugari sor daitezke. Arrisku hauek proiektuaren arrakasta baldintzatu dezakete. Horregatik garrantzitsua da arrisku hauek garaiz identifikatzea. Horrela arriskuen ondorioak murriztu edo ekiditea lor daiteke. Jarraian proiekturako identifikatu diren arriskuak aurkeztu eta bakoitzarentzat kontingentzia plana zehazten da.

1.5.1 Identifikaturiko arriskuak

1. Proiektuaren planifikazioa betetzeko gai ez izatea denbora arazoengatik, irakasgaien lan karga edo proiektuaren tamaina dela medio.

2. Proiektuaren parte diren datu guztien edo zatiren baten galera.

1.5.2 Kontingentzia plana

1. Arazo hau ekiditeko planifikazio malgu bat diseinatu da, horrela irakasgaiek lan karga handia duten momentutan proiektuaren lan karga atzeratzeko aukera izango da. Hala ere azken kasu batean proiektua 2 hilabete atzeratzeko aukera existitzen da, irailean beste deialdi bat baitago eta bertan proiektua defendatzeak ez bailuke ikaslearen ikasketetan atzerapenik suposatuko.

2. Arrisku hau murrizteko astero proiektu osoaren segurtasun kopiak gordeko dira, bestalde lana *dropbox* direktorio batean egingo da, horrela edozein disko arazo gertatzen bada beti egongo dira datuak atzigarri lainoan.

2 Kapitulua

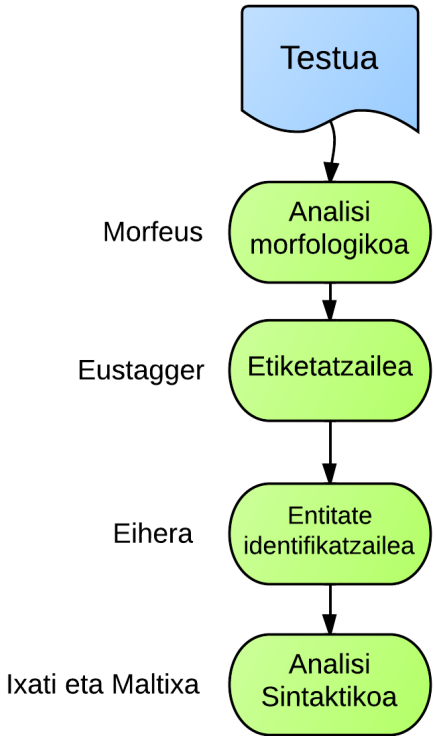
Hizkuntzaren prozesamendurako tresnak

Gaien Aurkibidea

Kapitulu honetan proiektua garatzeko erabili diren hizkuntzaren prozesamendurako tresnak aurkeztuko dira. Hala nola, erabilitako aplikazioak, formatuak, baliabideak eta corpusak.

2.1 Aplikazioak

Atal honetan hizkuntzaren prozesamendurako erabilitako aplikazioak azalduko dira. Lehenik testutik hurbilen jarduten duten aplikazioei buruz hitz egingo da, pixkanaka hauek konbinatuz analizatzaile sintaktikoak azaltzera iritsi arte. 2.1 irudian ikusiko diren programen laburpen bat erakusten da.



Irudia 2.1: Analizatzaileen laburpena

2.1.1 Analisi morfologikoa

Analizatzaile morfologikoa[1] azalduko diren aplikazioetatik testutik hurbilen lan egiten duena da. Honek testuko hitzak banan-banan hartu eta morfologikoki hitz horri dagozkion aukera guztiak itzultzen ditu. Honetarako egoera

[1] Analisi morfologikoa burutzeko Ixa ikerketa taldeko [2] *morfeus* aplikazioa erabili da.

finituzko transduktore bat erabiltzen du. Bere emaitza hitz bakoitzaren morfema zerrenda bat da, hau da, lema, pluraltasun, mugatasun eta bestelako markak. Honi dagokion beste zenbait informazio ere gehitzen dio, hala nola, sintagmak izan dezakeen kasua eta esaldian izan dezakeen kategoria(aditza, izena, adjektiboa...). Hona hemen adibide bat:[2]

Txoriak	**hori**	**dakar**
txori+ak(IZEARR+ABS)	horitu+0 (ADISIN+AMM)	dakar (ADT)
txori+ak(IZEARR+ERG)	hori (ADJARR)	
	hori+0 (DETERK+ABS)	

Taula 2.1: Analisi morfologikoaren adibide bat

Adibidean *Txoriak hori dakar.* esaldiaren analisi morfologikoa erakusten da. Bertan *txoriak* hitzak bi analisi ezberdin izan ditzakeela ikus daiteke. Bietan bere kategoria izen arrunta izango da, baina batean kasua absolutiboa izango da eta bestean ergatiboa. *hori* hitzak 3 aukera aurkezten ditu, aditza, adjektiboa edo determinatzailea izatekoa. Azkenik *dakar* hitzak aditz trinkoa izateko aukera soila du.

Hitzen kategoria jakitea garrantzitsua izango da gure zereginetarako. Termino erauzleak adibidez, *izena* kategoria duten hitzak identifikatuko ditu terminotzat eta galdera sortzaileak galdera baten erantzunean ez du inoiz aditz soil bat utziko. Bestalde, kasuaren informazioa oso garrantzitsua izango da galdetzaileak determinatzeko orduan, kasu askok galdetzaile bat edo batzuei erantzuteko joera baitute askotan.

2.1.2 Etiketzailea

Etiketatzailetzat (*Part of Speech tagger*) *Eustagger* aplikazioa erabili da, honek 2.1.1 atalean azaltzen den *morfeus* analizatzaile morfologikoa du integratuta. *Eustagger*en helburua analizatzaile morfologikoak ematen dituen analisi posible guztien artean, esaldi osoa kontuan hartuta, probabilitate handiena duena aukeratzea da. Hau horrela, 2.1 ataleko analisia honela murriztuko luke:

[2]Kasu batzuk kendu egin dira taulatik, hau garbiago egiteko.

Txoriak	hori	dakar
txori+ak(IZEARR+ERG)	hori+0 (DETERK+ABS)	dakar (ADT)

Taula 2.2: Analisi morfologikoa, etiketatzaile ondoren

2.1.3 Entitate izendatu identifikatzailea

Entitate izendatu identifikatzaile (*Named Entity Recognizer*)[3] batek testu bati buruzko informazio semantikoa ematen digu. Bere helburua hitz edo hitz multzo batek pertsona, leku edo instituzio bat errepresentatzen duen adieraztea da. Adibidez *Etxe Zuria* entitate izendatu bezala etiketatuko du Ameriketako Estatu Batuetako presidentearen etxeari buruz ari bagara, baina ez edozein etxe zuri arrunti buruz ari bagara. Hau detektatzeko datubasean gordetako entitateak zein datu meatzaritza tekniketatik lorturikoak erabiltzen ditu.

Galdera sorkuntzan informazio hau baliagarria gerta daiteke galdetzaileen desanbiguazioa burutzeko. Erantzuna leku izen berezi bat dela baldin badakigu, galdetzailea lekuzko galdetzaile bat izango dela jakin baitezakegu.

2.1.4 Analizatzaile sintaktikoak

Analizatzaile sintaktiko baten helburua esaldi baten sintagma zein berauen arteko dependentziak detektatzea da. Hau egiteko bi analizatzaile ezberdin erabili dira. Bat azaleko sintaxia (Ixati) burutzeko eta bestea sintaxi sakonago (Maltixa) bat burutzeko. Bi analizatzaileok aurretik aipaturiko erreminta guztiak dituzte integratuta, hau da, esaldi baten analisi sintaktikoaz gain, bere informazio morfologiko zein entitate izendatuei dagokiona ere aurkezten digute.

2.1.4.1 Sintaxi partziala[8]

Sintaxi partzialaren helburua ez da analisi guztiz zehatz bat sortzea. Bere helburua esaldi bateko *chunk*ak detektatzea da. *Chunk* bat, bata bestearen alboan dauden eta elkarrekin erlazioa duten hitz multzo bat da. Honek sintagma bat osa dezake edo ez. *Chunk*ak detektatzeaz gain hauek bitan sailkatzean ditu, aditz sintagma zein izen sintagma bat osa dezaketenak. Hona

[3]Proiektuan erabili den entitate izendatu identifikatzaileak *Ehiera* du izena eta Ixa ikerketa taldean garatutako da.

hemen adibide bat:

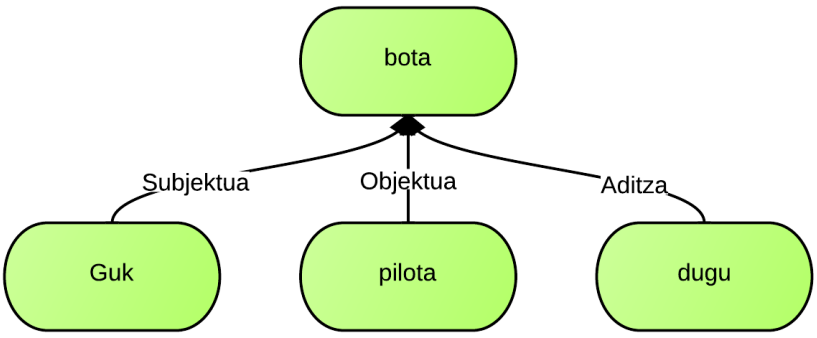
[Hartutako neurriek] [emaitza onak] [eman dituzte].

Izen sintagmak berdez, eta aditz sintagmak gorriz. Proiektuan azaleko sintaxia burutzeko Ixa ikerketa taldeko *Ixati*[8] aplikazioa erabili da. Bere irteera 2.3.1 atalean azaltzen den *Kyoto Annotation Framework(KAF)* formatuan adierazten da.

Sisteman aplikazio hau termino erauzlean moduluan erabiltzen da, bertan ez baita analisi sintaktiko sakon baten beharrik.

2.1.4.2 Sintaxi osoa[19]

Sintaxi osoko analizatzaileen helburua esaldi baten zuhaitz itxurako errepresentazio bat sortzea da. Zuhaitz egitura horretan nodoak hitzak dira eta ertzak sintagmen arteko dependentziak. Sakoneko analisia burutzeko Ixa ikerketa taldeko *Maltixa* [19] aplikazioa erabili da. Bere irteera 2.3.2 atalean azaltzen den *Computational Natural Language Learning(CoNLL)* formatuan adierazten da. Hona hemen adibide bat:



Irudia 2.2: Dependentziak

Dependentzia hauek oso interesgarriak dira esaldian transformazioak burutzeko eta galdera sortzaile moduluan horretarako erabiltzen dira.

2.2 Baliabideak

Jarraian proiektuan erabili diren baliabideen inguruko azalpenak ematen dira. Lehenik rol semantikoen hiztegiari buruz hitz egingo da. Ondoren bizi-

```
1 abestu ERG=agent=Nork
2 afaldu INE=TMP=Noiz
3 agertu ABS=Theme=Zer INE=LOC=Non
```

Irudia 2.3: Rol semantikoen adibidea

dun/bizigabe hiztegiari buruz. Amaitzeko WordNet zer den azalduko da.

2.2.1 Rol semantikoen hiztegia

Rol semantikoen hiztegiak [11] maiz agertzen diren 113 aditzen inguruko informazioa biltzen du. Aditz bakoitzarentzat, bere esaldian ager daitezkeen deklinabide kasuentzat aukera gehien duen rol semantikoa zehazten da.

2.3 irudian ikus daiteken moduan *afaldu* aditzaren esaldian dauden inesibo kasuek denborazko zentzu semantikoa izateko probabilitate handia dute. Hau jakinda *Noiz* galderari erantzuten diotela ondoriozta daiteke. *agertu* aditzean ere inesibo kasua maiz agertzen da, baina kasu honetan inesibo kasudun hitz multzoek lekuzko zentzu semantiko bat izateko aukera gehiago dute. Horrela *agertu* aditza dagoen esaldietan agertzen diren inesibo kasuek *Non* galdetzaileari erantzuteko probabilitate handia dute.

2.2.2 Bizidun/Bizigabe hiztegiak

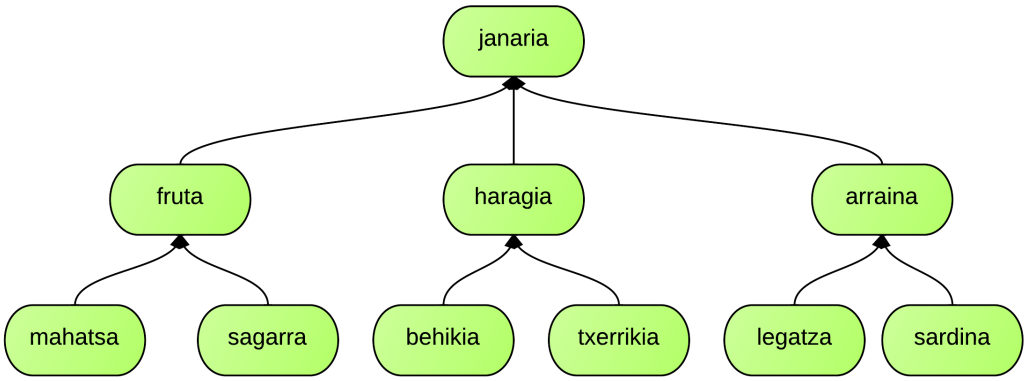
Bizidun/bizigabe hiztegia [14] Ixa taldeko itzulpengintza arloan erabiltzen den hiztegi semantiko bat da. Bertan 15.000 izen baino gehiago daude *bizidun*, *bizigabe* edo *zehazgaitz* bezala etiketatuak.

Sisteman informazio hau ere oso garrantzizkoa da galdetzaileak desanbiguatzeari begira.

2.2.3 WordNet

WordNet [18] hizkuntza ugaritan aurki daitekeen datu-base lexikal handi bat da. Bertan izenak, adjektiboak, aditzak etab. sinonimia, hiperonimia, hiponimia eta beste zenbait erlazioen arabera antolatuta aurkitzen dira. Egitura hau oso interesgarria da hizkuntzaren prozesamendua burutzeko. 2.4 irudian bere egituraren adibide bat ikus daiteke. Bertan nodo gorena *janari* kontzeptua da, bere hiponimoak *fruta*, *haragi* eta *arrain* direlarik. Hauek, aldi

berean, bere hiponimoak dituzte. Horrela behera eta goraka jarraituz grafo erraldoi bat dago sortuta *WordNet* datu-basean.



Irudia 2.4: Wordneten egituraren adibide bat

Sisteman informazio hau adieretan oinarrituriko erauzketa burutzeko erabiltzen da. Adiera berdina errepresentatzen duten terminoak erlazionatzeak termino erauzketa modu inteligenteago batean burutzeko aukera ematen du aurrerago ikusiko den moduan.

2.3 Formatuak

Jarrian proiektuan erabili diren bi formaturen berri ematen da. Lehenik *Kyoto Annotation Framework (KAF)* formatuari buruz hitz egingo da eta ondoren, *Computational Natural Language Learning(CoNLL)* formatuari buruz.

2.3.1 Kyoto Annotation Framework (KAF)

Kyoto Annotation Framework (KAF) [6] testuak linguistikoki etiketatzeko formatu bat da. *XML*n oinarritua dago eta bertan testu baten analisia gordetzen da, maila morfologiko, sintaktiko eta semantikoan. Gure sisteman formatu hau 2.1.4.1 atalean azaltzen den Ixati analizatzailearen irteera formatu moduan erabiltzen da. Honako xehetasunak aurkezten ditu: *XML* formatua jarraitzen duenez lehenik *XML* goiburukoa dator, bertan *XML* bertsioa eta kodeketa mota zehazten direlarik. Honen jarraiki dokumentuaren erro nodoa irekitzen da, *KAF* bezala izendatua.

```
1 <?xml version="1.0" encoding="UTF-8" standalone="no"?>
2 <KAF xml:lang="eu">
```

Lehen datu bezala testuko *wordForm*ak datoz, hau da, hitzak testuan agertzen diren moduan. Hauek *wf* etiketarekin zehazten dira. *WordForm* bakoitzean berau zenbatgarren esalditik lortutako den adierazten da *sent* atributuaren bitartez.

```
<text>
<wf wid="w1" sent="1">Gaur</wf>
<wf wid="w2" sent="1">eguraldi</wf>
<wf wid="w3" sent="1">ona</wf>
<wf wid="w4" sent="1">dago</wf>
<wf wid="w5" sent="1">.</wf>
</text>
```

Hurrengo nodo mota *term* motakoa da. Honek termino bat sinbolizatzen duelarik. Termino bat hitz bat bakarrik edo hitz anitzeko unitate bat izan daiteke, adibidez *Estatu Batuak* hitz bat baino gehiagoko termino bat izango litzateke. Termino bakoitzean aurkezten den informazioari dagokionez bere lema eta eta kategoria (*pos*) agertzen dira. Lema hitz baten forma normala da, hau da, hiztegi batean aurki dezakegun forma. Kategoria, aldiz, hitzak esaldiaren testuinguruan duen funtzioa da, hala nola, aditza, izena, adjektiboa... Termino batzuetan bere sintagmaren kasua ere adierazten da *case* atributuaren bidez.

```
<terms>
<!-- Gaur -->
<term tid="t1" type="open" lemma="gaur"
   pos="A.ADB-ARR">
<span>
<target id="w1"/>
</span>
</term>
<!-- eguraldi -->
<term tid="t2" type="open" lemma="eguraldi"
   pos="N.IZE-ARR">
<span>
<target id="w2"/>
</span>
</term>
<!-- ona -->
<term tid="t3" type="open" lemma="on" pos="G.ADJ-ARR"
   case="ABS">
```

```
<span>
<target id="w3"/>
</span>
</term>
<!-- dago -->
<term tid="t4" type="open" lemma="egon" pos="V.ADT">
<span>
<target id="w4"/>
</span>
</term>
</terms>
```

Azken nodo mota bezala *chunk*-ak ditugu. *Chunk* bat termino multzo bat da, esaldian funtzio sintaktiko bat izan dezakeena. *Chunk* analisia nolabaiteko analisi sintaktiko partzial bat da. *Chunk* bakoitzean aditz edo izen sintagma bat den erakusten zaigu *phrase* atributuaren bidez. Hemen ere chunk batzuetan bere kasua erakusten zaigu, adibidez *eguraldi ona* chunka absolutibo kasukoa dela adierazten da.

```

<chunks>
<!-- Gaur -->
<chunk cid="c1" head="t1" phrase="NP">
<span>
<target id="t1"/>
</span>
</chunk>
<!-- eguraldi ona -->
<chunk cid="c2" head="t2" phrase="NP" case="ABS">
<span>
<target id="t2"/>
<target id="t3"/>
</span>
</chunk>
<!-- dago -->
<chunk cid="c3" head="t4" phrase="VP">
<span>
<target id="t4"/>
</span>
</chunk>
</chunks>
```

2.3.2 Computational Natural Language Learning(CoNLL)

Computational Natural Language Learning(CoNLL) formatua 2.1.4.2 atalean azaltzen den *Maltixa* analizatzaile sintaktikoaren irteera formatua da. Formatu hau taula formatu bat da, tabulatzailez banaturiko zutabez osatua eta lerro saltoz banaturiko errenkadez. Errenkada bakoitzean termino baten inguruko informazioa erakusten da. Taulak zuhaitz egitura bat errepresentatzen du, egitura hau *id* eta *burua* atributuen bitartez lortzen da. Maltixaren irteera baten adibidea 2.3 taulan eta 2.2 irudian ikus daiteke, taula moduan zein zuhaitz errepresentazioan hurrenez hurren.

id	Hitza	Burua	Dep.	Lema	Ezaug.	Kat.	Azpikat.
1	Guk	3	ncsubj	gu	KAS:ERG	IOR	PERARR
2	pilota	3	ncobj	pilota	KAS:ABS	IZE	ARR
3	bota	0	ROOT	bota	ADM:PART	ADI	SIN
4	dugu	3	auxmod	edun	_	ADL	ADL
5	.	4	PUNC	.	_	PUNT	PUNT

Taula 2.3: Maltixaren irteera

Honako zutabeak aurki ditzakegu bertan:

- **Id**: Hitzaren identifikatzailea
- **Hitza**: Hitza esaldian zetorren bezala.
- **Burua**: Hitzaren burua, zuhaitz egituran gurasoa izango litzatekeena.
- **Dependentzia**: Uneko hitzak bere buruarekiko (gurasoa) duen dependentzia sintaktikoa, subjektua, objektua...
- **Lema**: Hitzaren lema erakusten du, hau da, hitza hiztegian agertzen den forma normalean.
- **Ezaugarriak**: Bertan ezaugarri ezberdinak aurki ditzakegu, kasua, numeroa, aspektua...
- **Kategoria**: Hitzaren kategoria adierazten du, aditza, izena, adjektiboa...
- **Azpikategoria**: Kategoria atributuaren zehaztapen bat da, izenen kasuan arrunta edo berezia ote den adierazten da bertan adibidez. Entitate izendatuen kasuan ere hemen adierazten da informazio hau.

3 Kapitulua

Corpus azterketa

Gaien Aurkibidea

Ikerketa lanarekin aurrera jarraitu ahal izateko, bai probak burutzeko zein modulu batzuk entrenatzeko, domeinuetako corpusak lortzeko beharra ikusi da. Horregatik, historia, teknologia zein domeinu orokorreko corpusak identifikatu, lortu, konparatu eta horien artean egokienak aukeratu dira. Jarraian aipaturiko azterketa jorratzen da.

3.1 Analisia

Ixa ikerketa taldean[2] existitzen diren corpusak identifikatzeaz gain sarean aurki daitezkeen baliabideak ere analizatu dira. Honakoak izan dira identifikaturiko baliabide posibleak:

Egunkaria corpusa

Egunkaria corpusa *Egunkaria* aldizkariak urteetan zehar argitaratutako testuen bilduma bat da. Eguneroko berrien inguruko informazioa argitaratu izanak domeinu orokorreko corpus moduan erabiltzeko aproposa bihurtzen du, bertan arlo askotariko testuak aurki baitaitezke. Corpus hau Ixa ikerketa taldeko zerbitzarietatik hartu da.[1]

ZT corpusa

ZT corpusa[5] zientzia eta teknologiako testu bilduma bat da. Bere testu guztiak 7 eremutan etiketatuta daude eta eremu horietako bat teknologia da. Beraz teknologia corpus bat eratzeko aproposa izan daitekeela aurreikusten da. Ixa ikerketa taldeko zerbitzarietan aurki daiteke.[2]

Lur hiztegi entziklopedikoa

Lur hiztegi Entziklopedikoa[3] sarean aurkitu daitekeen baliabide bat da. Bertan jakintza arlo askoren informazioa kontsulta daiteke. Gai hauen artean historia unibertsalaren gaia dago, eta historia domeinuko corpus bat osatzeko aproposa izan daiteke.

Wikipedia

Wikipedia[1] erlazionaturiko dokumentu multzo handi batez osatua dago. Erlazio hauek erabiliz bai teknologia eta bai historiako testu bilduma batzuk lortu daitezkeela aurreikusten da.

[1] "jirxuxen/CORPORA/elebakarrak/egunkaria" helbidean aurki daiteke.
[2] zehazki honako helbidean: "/sc01a5/hizking/IXA/ZT_ling_morfo".

3.2 Eskurapena

Jarraian corpus bakoitza nola eskuratu den azaltzen da. Prozesu honetan ez da eskuraketa hutsa burutu, corpusak lortzeaz gain guztiak gerorako komenigarria izango den formatuan uzteko esfortzu bat ere burutu da. Honez gain corpus guztiak antzekoak izateko helburuarekin guztiak lur corpusaren(txikiena) tamainara murriztu dira, 200.000 hitz ingurura.

3.2.1 Egunkaria corpusa

Egunkaria corpusa Ixa ikerketa taldeko zerbitzarietan aurkitzen da. Corpusa testu lauan aurkitzen da, beraz ez da formatu aldaketarik burutu behar izan. Aldiz, bere tamaina 200.000 hitzetik gorakoa denez bere zati bat bakarrik hartu da. Zatia hartzean corpusaren izaera orekatua mantendu nahi izan da. Corpusa domeinu orokorrekoa da bere osotasunean, baina ez du zertan hala izan behar zati bat hartzen badugu. Adibidez, egun guztietako lehen 15 dokumentuak hartzen baditugu, domeinu orokorra izatetik, ekonomia domeinukoa izatera pasa daiteke, egunero lehen orrietan ekonomiaz hitz egiten baita gehien. Horregatik zatia erauzteko moduak egunak bere osotasunean errespetatu ditu. Guztira bi hilabete oso hartu dira 200.000 hitz inguru osatzeko.

3.2.2 ZT corpusa

ZT corpusa[5] zientzia eta teknologia arloan espezializaturiko corpus bat da. Berau Ixa ikerketa taldeko zerbitzarietan aurkitzen da *XML* eta *TEI* formatuak erabiliz. Gure interesekoa ordea, teknologia azpidomeinua da. Testuen *TEI* metainformazioaren artean testu bakoitzaren domeinuari buruzko informazioa agertzen da. Etiketa horretaz baliatu gara teknologia azpidomeinukoak diren testuak lortzeko. Hauek lortuta oraindik corpusa handiegia da eta ausaz aukeratu dira testu batzuk 200.000 hitz inguru osatzeko. Testuei *XML* eta *TEI* formatua kendu zaizkie, hauek testu lau bihurtuz, aplikazioaren funtzionamenduari begira.

3.2.3 Lur hiztegi entziklopedikoa

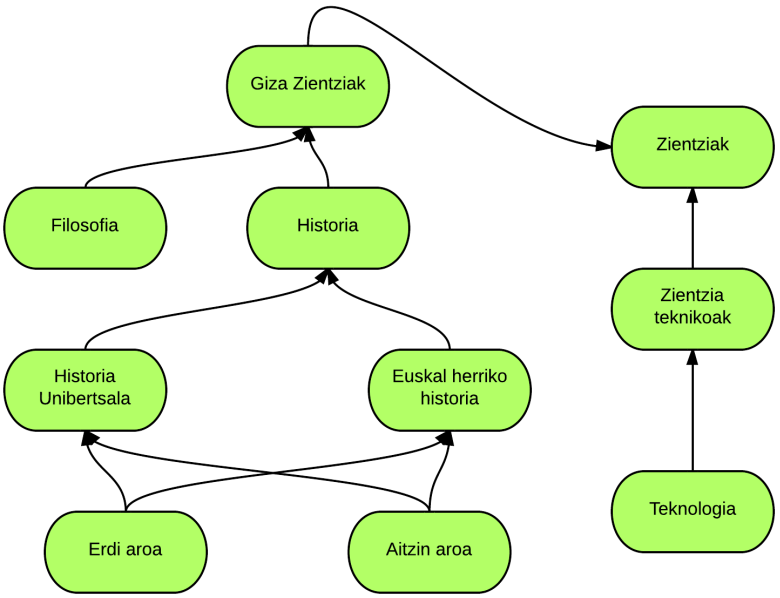
Lur hiztegi Entziklopedikoa[3] sarean aurki daitekeen euskarazko entziklopedia bat da. Hiztegiaz gain jakintza arlotan ordenaturiko atal bat ere badu, zenbait ezagutza arlo jorratzen direlarik bertan. Ezagutza arlo hauetariko bat historia unibertsala da. Bertako testuak armiarma (*crawler*) baten la-

guntzaz bildu dira. Bertatik 200.000 hitz inguruko corpus bat bildu da eta testu lauan utzi da prestatuta.

3.2.4 Wikipedia

Wikipedia [1] sarean aurki daitekeen elkarren artean erlazionaturiko testu bilduma bat da. Erlazio hauek guztiek grafo egitura bat osatzen dute, bertako nodoak, kategoria nodoak zein dokumentu nodoak izan daitezkeelarik. Grafo honen adibide bat ikus daiteke 3.1 irudian.

Grafo honetaz baliatuz, kategoria bat eman eta berarekin zuzenki edo zeharka erlazionatuta dauden dokumentu nodo guztiak lor daitezke eta hau egiten duen programa bat sortu da.[3] Behin programa izanda, bai teknologia, bai historiako domeinuko testuak erauzi dira. Hemen ere testuak ausaz lortu dira 200.000 inguru hitz lortzeko. Corpus hau ere testu lauan utzi da.



Irudia 3.1: Wikipedia grafoaren adibidea

3.3 Konparaketa eta aukeraketa

Puntu honetara iritsita, behar genituen 3 domeinutako corpusak ditugu, guztiak 200.000 hitz ingurukoak eta testu lauan gordeta. Guztira 5 corpus ditu-

[3] "/sc01a4/users/jmadrazo003/Lanak/Wikipedia_gaika/wminer_eu_20121112/wikipedia_gaika" helbidean aurki daiteke.

gu:

- Domeinu orokorreko 1, *Egunkariako*a
- Historiako 2, bata *Wikipedia*koa eta bestea *Lur Hiztegi Entziklopediko*koa
- Teknologiako 2, bata *Wikipedia*koa eta beste *ZT corpus*ekoa

Beraz dudarik ez dago domeinu orokorreko corpus bezala *Egunkaria* corpusa erabili beharko dela, hauxe baita domeinu orokorrean dugun bakarra. Beste bi domeinutarako eskuzko berrikusketa bat burutu da eta domeinu bakoitzerako corpus bana aukeratu da. Eskuzko berrikusketaren ondorioz Wikipedian itxarotakoa baino zarata handiago zegoela ikusi da. Bertan zerrenda, titulu eta esteka kopuru uste baino handiagoa dago, eta honek gure zereginetako zailtasunak handiegiak ekar ditzake epe laburrean konpondu ahal izateko. Aldiz, beste bi corpusetan (ZT eta Lur) testuak garbiagoak dira, ez dute ez esteka eta ez zerrendarik normalean, eta testuen arteko tamainak ere orekatuagoak dira Wikipedian baino. Arrazoi hauek guztiak direla medio, Wikipediatik bilduriko corpusak albo batera uztea erabaki da.

4 Kapitulua

Termino Erauzlea

Gaien Aurkibidea

Atal honetan, proiektuan gauzatu diren bi ekarpenetatik lehenengoa azaltzen da, hau da, termino erauzleen inguruan burutu den azterketa. Aurrekari azterketa bat burutzeaz gain, bi sistema ezberdinen euskararako analisi, diseinu eta inplementazioa aurkezten dira. Amaitzeko bi sistemen zein euskararako existitzen den beste sistema baten arteko konparaketa bat burutzen da.

4.1 Aurrekari azterketa

Eduki aukeraketa burutzeko teknika ezberdinak landu izan dira. Bi teknika ezberdin aztertu dira sistemarako. Batetik kontzeptu mapetan oinarrituriko eduki aukeraketa eta bestetik termino erauzleetan oinarriturikoa. Mapa kontzeptualen inguruko eduki aukeraketaren inguruko azterketa 5.1.1.1 atalean ikus daiteke.[1] Proiektuaren denbora mugen ondorioz termino erauzleetan oinarrituriko eduki aukeraketa zentratzea erabaki da.

Termino erauzle bat testu batetik domeinu batekiko esanguratsuak diren hitz edo hitz multzoak erauzteko tresna linguistiko bat da. Atal honetan domeinuko termino erauzle sistemen inguruan existitzen diren aurrekariak aurkeztuko dira. Lehenik euskara ez diren beste hizkuntza batzuetarako sistemak aurkeztu eta modu laburrean azalduko dira, geroago sakonago azalduko baitira. Ondoren bereziki euskarako dagoena aztertuko da.

4.1.1 Beste hizkuntzetarako aurrekari azterketa

Azken urteetan, terminoen erauzketa teknikekiko interesa handitzen ari da, beste zenbait lanetan izan dezaketen esangura dela medio. Horregatik, terminoen erauzpenerako tresna ugari garatu izan dira, hala nola, *Alegria I. et al.*-ek (2004) aipatzen dituzten ACABIT, ANA, LEXTER, TERMINO, TERMS, Xtract, CLARIT, FASTR, NODALIDA...

Proiekturako bi termino erauzte teknika jo dira interesgarritzat, beraien emaitza zein erabiltzen dituzten teknika bereziengatik. Biek teknikek estatistikoak erabiltzen dituzte beren helburua lortzeko. Jarraian modu laburrean aurkezten dira, 4.3 eta 4.4 atalean sakonago azalduko baitira.

[1] Galdera eraikuntzarako metodo baliagarria ere bada kontzeptu mapena, hortik, atal horretan agertzea.

4.1.1.1 Ontologiak sortzeko termino erauzketa[26]

Velardi et al.-ek (2001) ontologiak sortzen laguntzeko *SymOntos* izeneko sistema bat proposatzen dute. Sistema honek termino erauzle bat du oinarritzat.

Sistema metodo estatistikoetan oinarritua dago eta metodo estatistiko gehienen moduan corpus batzuetatik lorturiko ikasketan oinarritzen du bere funtzionamendua. Behin corpusaren gainean ikasketa burutua duelarik, testu berrietan terminoak detektatzeko gai da, sorturiko datu-basean bilaketak burutuz.

Ikasketa prozesua burutzeko, bi estatistiko ezberdin erabiltzen dira sisteman.

Batetik, teknika hauetan arrunta den *tf.idf (term frequency* $\times$ *inversed document frecuency)* antzeko estatistiko bat aplikatzen da, *relebantzia* bezala izendatua. *tf.idf* baten funtzionamendurako bi corpus bereizten dira, domeinukoa eta orokorra. Estatistikoak terminoaren maiztasunak kalkulatzen ditu bi corpusetan eta bi hauen konparaketa bat burutzen du. Termino bat domeinu corpusean, corpus orokorrean baino gehiagotan agertzen bada, termino hori domeinuarekiko esanguratsua dela esan daiteke. Aldiz, terminoa domeinu corpusean, corpus orokorrean bezalako maiztasun edo txikiagoan agertzen bada terminoa ez dela domeinukoa esan daiteke.

Bestetik, berezitasun bezala, entropian oinarrituriko bigarren estatistiko bat erabiltzen du. Honi *kontsentsua* deitzen diote, eta terminoa corpusean zehar nola banatuta dagoen sinbolizatzen du. Termino bat zenbat eta modu uniformeagoan banatuta egon corpusean orduan eta ziurragoa da terminoaren esangura. Aldiz, termino bat corpuseko dokumentu bakar batean agertzen bada, termino hau zarata besterik ez izateko aukera handia dago.

4.1.1.2 Adieretan oinarrituriko termino erauzketa[24]

Scaleanu B. et al.-ek (2001) aurrera pausu kualitatibo bat ematen dute termino erauzketan. Tresna hau ere metodo estatistikoetan oinarritzen da, *tf.idf* estiloko teknika batean hau ere.

Baina, berezitasun bezala, bertan hitzen maiztasunak erabili beharrean kontzeptuen maiztasunak erabiltzen dituzte. Horrela adibidez, janari, jaki

eta elikagai hitzak ez lituzke independenteki neurtuko, kontzeptu bat bezala baizik. Testu askotan hitz bera ez errepikatzearren sinonimoak erabiltzeko joera izaten da, sistema honek sinonimoak detektatzen ditu, testuko kontzeptu garrantzitsuak detektatu ahal izateko. Sinonimo hauek detektatzeko 2.2.3 atalean aurkezten den *WordNet* zerbitzua erabiltzen da.

4.1.2 Euskararako aurrekari azterketa

Euskararako termino erauzleen inguruan garatu den ikerketa murritzagoa da beste hizkuntzetarako garatutakoa baino. Arlo honetako aurrekarien artean aipatzekoak Tybot zein Erauzterm sistemak dira. Bere emaitza hobeak direla medio, Erauzterm sistema hartu da proiektuko erreferentziatzat eta berau sakonki aztertua izan da.

4.1.2.1 Erauzterm

Erauzterm[17] euskararako termino erauzle bat da. Bere helburua lortzeko informazio linguistiko zein estatistikoa erabiltzen du. Berezitasun bezala, terminoak erauzteko bi teknika ezberdin erabiltzen ditu, bat hitz bakarreko terminoentzat eta bestea hitz anitzekoentzat.

Hitz bakarreko terminoentzat *tf.idf* moduko teknika bat erabiltzen du, hau da, terminoen maiztasunen konparaketa bat egiten du bi corpusen gainean.

Hitz anitzeko terminoei dagokienez, terminoa osatzen duten hitzen arteko informazio kantitatea neurtzen da. Honetarako *mutual information*, *log-likelihood*, *mutual expectation* eta *chi-square* estatistikoak erabiltzen dira. Estatistiko guzti hauek hitz bat beste baten alboan agertzeko probabilitateak neurtzen dituzte corpus orokor batean eta domeinuko corpusean fenomeno hau gertatzeko maiztasunekin konparatzen dute. *tf.idf* tekniketan ez bezala, teknika hauetan probabilitateak hitzaren inguruan dauden beste hitzek baldintzatuak dira.

4.2 Zer garatuko da

Aurrekariak aztertu ondoren, metodo berri bat sortzea baino, hauetan ikusi diren metodoak euskararako moldatzea interesgarriagotzat jo da. Horregatik, bai *Scaleanu B. et al.*-ek (2001) azaldutako metodoa (aurrerantzean TE1 deitua) eta bai *Velardi et al.*-ek (2001) azaldutako metodoa (aurrerantzean

TE2 deitua) euskararako inplementatzea erabaki da. Kontzeptu mapen inguruan aztertu ziren sistemak baztertuak izan dira, hauek inplementatzeak denbora aldetik kostu handia suposatuko bailuke.

4.3 Adieretan oinarrituriko termino erauzlea (TE1)

Jarraian adieretan oinarrituriko termino erauzlea (TE1) garatzeko egin beharrekoak aurkezten dira. Honetarako lehenik metodoa deskribatuko da. Ondoren metodoa inplementatzeko, diseinu aldetik hartu diren erabakiak azalduko dira.

4.3.1 Metodoaren deskribapena

Jarraian adieretan oinarrituriko termino erauzlearen metodoaren deskribapen bat azaltzen da. Bertan berau osatzen duten estrategia ezberdinen inguruan hitz egingo da.

4.3.1.1 Aurreprozesua

Metodoaren lehen pausuak testu lauan aurkitzen diren corpusak aurreprozesatzean datza. Testuan gertatzen diren zenbait fenomeno tratatzea beharrezkoa baita.

Batetik, termino batek ez du zertan beti hitz bakar bat izan behar, batzuetan HAUL (hitz anitzeko unitate lexiko) bat ere izan daiteke. Adibidez, *eremu magnetiko* termino bezala hartzea nahi da, eta ez bi termino ezberdin bezala.

Bestetik, beharrezkoa da terminoak nolabait normalizatzea. Adibidez, *etxe* terminoa modu askotan ager daiteke corpusean: *etxearekin*, *etxetik*, *etxera*... Terminoaren forma guzti hauek tratatu behar dira, guztiak bat direla identifikatzeko. Fenomeno hau tratatzeari lematizazioa deitzen zaio. Lema hitz baten forma normala da, hiztegian agertzen dena. Testuan agertzen den hitz bakoitzarentzat bere lema lortu nahi da.

Azkenik, garrantzitsua da hitz bakoitzaren kategoria jakitea. Termino batek ezin izango baitu adibidez aditz kategoriakoa izan. Termino guztiek izen soilak izan beharko dute. Hau jakiteko testua etiketatzaile (Part of speech

tagger) batekin prozesatu beharko da.

Aurreprozesua burututakoan, testu lautik termino hautagaiak lortuko dira. Hauek hitz bakarreko zein hitz anitzekoak izan ahalko dira, baina beti ere izenak izan beharko dute. Termino hautagai hauek hasieran zuten forma albo batera utzi eta beren lema moduan aurkeztuko dira hurrengo prozesuak ongi burutzeko.

4.3.1.2 Terminoen relebantzia neurtzen

Hurrengo pausoa, termino bakoitzaren relebantzia neurtzea da. Hau egiteko *tf.idf*ren bertsio moldatu bat erabiltzen da. Honetarako jarraian erakusten den ekuazioa aplikatzen da.

$$rlv(t|d) = log(tf_{t,d})log(\frac{N}{df_t}) \tag{4.1}$$

Bertan t terminoa da, d domeinua eta N erabiliko diren domeinu kopurua (gure kasuan 3). Formula honek batetik terminoaren domeinuko maiztasuna kalkulatzen du: $tf_{t,d}$. Pisu hau formularen bigarren zatiaren bitartez $log(\frac{N}{df_t})$ doitzen da. df_t terminoaren domeinu bateko agerpen kopurua izanik, domeinu guztietan agertzen den termino bati pisu osoa emango dio. Aldiz, domeinu bakarrean agertzen diren terminoei pisu osoa emango zaie, eta zenbat eta domeinu gehiagotan agertu pisu txikiagoa emango zaie.

4.3.1.3 Kontzeptuen relebantzia neurtzen

Behin termino bakoitzaren relebantzia neurtuta kontzeptuen relebantziara pasatzea da hurrengo pausoa. Honetarako 2.2.3 atalean aurkezten den Wordnet zerbitzua erabiliko da. Bertan terminoak *Synset*etan ordenaturik daude. *Synset* batek kontzeptu bat errepresentatzen du, eta bertan kontzeptu bera errepresentatzeko posible diren termino guztiak daude sartuta. Hau horrela, kontzeptu baten relebantzia bere barneko termino guztien relebantzien batura bezala neur daiteke. Jarraian agertzen den formulan c kontzeptu baten pisua nola neurtu ikus daiteke.

$$rlv(c|d) = \sum_{t \in c} rlv(t|d) \tag{4.2}$$

Estaldura lexikala
Aurkeztutako formulak, ordea, nahi ez diren portaera batzuk aurkezten ditu

kasu batzuetan. Demagun hurrengo kasua:

Adiera 1: (cell, prison cell) Adiera 2: (cell)

Bertan bi *Synset* agertzen dira. Lehena kartzela gela kontzeptua errepresentatzeko eta bigarrena zelula. Testuan *cell* hitza bakarrik agertuko balitz, bi *Synset*ei pisu bera emango litzaieke, eta hori ez da nahi dena. Hau ekiditeko estaldura lexikala deituriko kontzeptu bat erabiliko da. Honekin *Synset*ean dauden hitzetatik zenbat agertu diren testuan kontuan izan nahi da, aurretik aipaturiko kasuak ekiditeko.

$$rlv(c|d) = \sum_{t \in c} \frac{T}{|c|} rlv(t|d) \tag{4.3}$$

Ekuazioan T *Synset*etik corpusean agertu den termino kopurua da eta $|c|$ *Synset*ean dagoen termino kopurua dira.

Hiponimoak
Estaldura lexikalak arazo batzuk konpontzen ditu baina oraindik beste batzuk mantentzen dira:

Batetik, termino bakarreko *Synset*ak hobesten ditu. Bertako terminoa agertzen denean pisu maximoa ematen baitie. Bestetik, pisu berdina ematen die hitz berdina duten luzera bereko *Synset*ei.

Hau konpontzen saiatzeko kontzeptu batean dagoen informazioa zabaltzea erabaki da. Hau horrela kontzeptuan sinonimoak kontuan hartzeaz gain terminoen hiponimoak ere kontuan hartuko dira. Honi kontzeptu zabaldua (c^+) deitzen diote. Hiponimo bat kontzeptu baten espezializazioa da, adibidez *animalia* hitzaren hiponimoak *katua* edo *txakurra* dira. Espezializazio honek termino bat benetan zein kontzeptutatik datorren zehazten laguntzen du. Jarraian agertzen den formularen arabera kalkulatzen da kontzeptu zabaldu (c^+) baten pisua.

$$rlv(c^+|d) = \sum_{t \in c^+} \frac{T}{|c|} rlv(t|d) \tag{4.4}$$

Ekuaziora erreparatuz gero, hiponimoak estaldura lexikalaren doiketan ez direla kontuan hartzen ikus daiteke. Honen arrazoia, hiponimo asko dituzten

kontzeptuak ez zigortzea da. Hala ere penalizazio txikiago bat ezartzea nahi da.

$$p = \frac{rlv(c|d)}{|h|} \tag{4.5}$$

Hau da, $|h|$ kontzeptuan agertzen den hiponimo kopurua izanik, kontzeptuaren relebantziari agertzen ez den hiponimo bakoitzeko p penalizazio bat ezarriko zaio. Penalizazioa kalkulatzeko moduak relebantzia inoiz ez dela 0 baino txikiagoa izango bermatzen du.

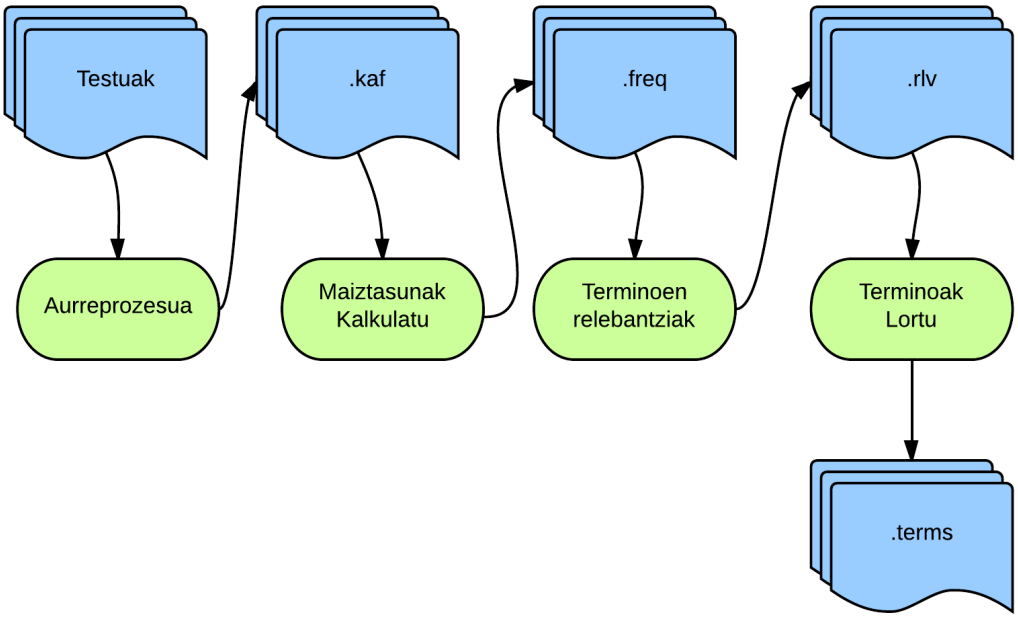
Azkenik, datuetan detektaturiko zenbait arazo ekiditeko, 30 hiponimo baino gehiago dituzten kontzeptuak baztertu egin dira. Horrenbeste hiponimo dituzten kontzeptuak askotan orokorregiak baitira. Adibidez *pertsona* kontzeptuak izan ditzakeen hiponimoak (*mutil*, *neska*, *aitona*, *amona*, *ama*,*aita*...) ugari dira.

4.3.2 Aplikazioaren diseinua

Jarraian termino erauzlearen inplementazioa burutzeko diseinu aldetik hartu diren erabakiak aipatzen dira.

4.3.2.1 Arkitektura

Corpusak oso handiak izan daitezke, eta hauekin lan egiteak denbora aldetik kostu handia suposa dezake. Horregatik, terminoak erauzteko programa modu banatu batean exekuta daitezkeen azpiprogrametan banatzea erabaki da. Azpiprograma horietako bakoitzak fitxategi batzuetatik irakurtzen ditu datuak eta bere emaitza ere fitxategi ezberdinetan utziko du. Horrela prozesu bakoitza modu independente batean exekutatzea ahalbidetzen, eta jada kalkulatuta dauden datuak berrerabiltzeko aukera ematen du. Programa nagusian honako azpi programak identifikatu dira:



Irudia 4.1: TE1en azpiprogramen egitura

Aurreprozesua(Komuna)
Prozesu hau komuna da bai TE1 eta bai TE2 termino erauzleetan. Bertan testuko hitzak lematizatu, HAUL-ak detektatu eta hauetatik izena kategoriakoak hartzen dira. Prozesu hau 2.1.4.1 atalean azaltzen den *Ixati* [8] aplikazioak burutzen du, eta bere irteera 2.3.1 atalean azaltzen den *Kyoto Annotation Framework*[6] formatuan itzultzen da.

Maiztasunak lortu(Komuna)
Prozesu hau komuna da bai TE1 eta bai TE2 termino erauzleetan. Honen helburua aurreprozesuak sorturiko KAF fitxategiak hartu eta bertako termino hautagai guztien kopuruak lortzea da. Pasatako corpus bakoitzeko ".freq"fitxategi bat sortuko du.

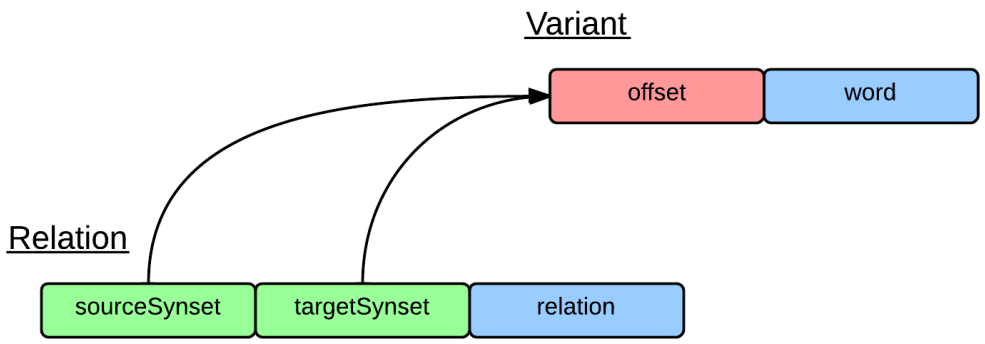
Relebantziak kalkulatu
Prozesu honek corpus bakoitzaren maiztasun fitxategiak hartu eta bertan agertzen diren termino bakoitzaren relebantziak neurtzen ditu 4.3.1.2 atalean azaltzen den formula aplikatuz. Emaitza ".rlv"luzapeneko fitxategietan gordetzen da eta corpus bakoitzarentzat emaitza ezberdina gordetzen du.

Terminoak lortu
Prozesu honek corpus bateko termino guztien relebantziak dituen ".rlv"fitxategia

hartu eta bere terminoak itzultzen ditu. Honetarako 4.3.1.3 atalean azaltzen diren kontzeptu relebantziak kalkulatzen ditu. Wordnet erabiliz *Synset*ak pisatzen ditu eta hauen pisuaren arabera ordenaturik itzultzen ditu bere terminoak. Emaitza ".terms"fitxategi batean gordetzen du.

4.3.2.2 Wordneten diseinua

Jarraian Wordnet zerbitzua gordeta dagoen datu-basearen zehaztapena erakusten da.



Irudia 4.2: Wordneten diseinua

Variant taula
Variant taula terminoak gordetzen dituen taula da eta ondorengo eremuak aurkezten ditu.

- **offset**
  Eremu hau identifikatzaile bat da, *Synset*aren barneko termino guztiek konpartitzen dute identifikatzaile berdina.

- **word**
  Eremu honetan terminoa gordetzen da, bere forma normalizatuan.

Relation taula
Bigarren taula honetan *WordNet* grafoko erlazioak gordetzen dira. Honako eremuak ditu:

- **sourceSynset**
  Eremu honek erlazioa hasten den *Synset*aren gakoa gordetzen du.

- **targetSynset**
  Eremu honek erlazioa bukatzen den *Synset*aren gakoa gordetzen du.

- **relation**
  Eremu honek erlazioaren mota gordetzen du, hiponimia, hiperonimia...

4.3.2.3 Wordnetekiko abstrakzio klaseak

*WordNet*en atzipena abstraitzeko helburuz klase batzuk sortu dira. Horrela programaren negozio logika eta modeloa banaturik mantenduko dira. Banaketa hau garrantzizkoa da kasu honetan, *Wordnet* proiektuaz kanpoko aplikazio bat baita. Horrela *WordNet*en aldaketa bat gertatuko balitz, ez litzateke termino erauzle osoa aldatu beharko, bakarrik honen atzipena burutzeko diseinaturiko klaseak baizik.

Abstrakzio hau burutzeko klase bat inplementatu da, *WordNetHelper* izenekoa. Klase honek *singleton* diseinu patroia betetzen du, hau da, programaren exekuzioa bere instantzia bakarra existituko da. Honako metodoak ditu:

- **connect**
  Datu basera konektatzeko balio du.

- **disconnect**
  Datu basetik deskonektatzeko balio du.

- **getTermsByOffset**
  Termino bat emanda bere *Synset* identifikadoreak itzultzen ditu.

- **getOffsetFromTerm**
  Synset identifikatzaile bat emanda bere termino guztiak lortzen ditu.

- **getHyponymsByOffset**
  Synset identifikatzaile bat emanda bere hiponimo guztiak lortzen ditu.

*Synset*ekin burutuko diren eragiketak errazteko asmoz *Synset* klase bat ere inplementatu da. Honek bere eragiketetan *WordNetHelper* klasearekin egiten du lan. Honako metodoak ditu:

- **eraikitzailea**
  Terminoa argumentuz jasota *Synset* egitura bat itzultzen du.

- **getTerms**
  *Synset*eko terminoak itzultzen ditu.

- **getHyponyms**
  *Synset*eko hiponimoak itzultzen ditu.

- **getRelebantzia**
  *Synset*aren relebantzia kalkulatu eta itzultzen du.

- **hiponimoGehiegiDitu**
  *Synset*ak 30 hiponimo baino gehiago ote dituen dio.

4.4 Kontsentsuan oinarrituriko termino erauzlea (TE2)

Jarraian kontsentsuan oinarrituriko termino erauzlea (TE2) garatzeko egin beharrekoak aurkezten dira. Honetarako lehenik metodoa deskribatuko da. Ondoren metodoa inplementatzeko, diseinu aldetik hartu diren erabakiak azalduko dira.

4.4.1 Metodoaren deskribapena

Jarraian kontsentsuan oinarrituriko termino erauzlearen metodoaren deskribapen bat azaltzen da. Bertan berau osatzen duten estrategia ezberdinen inguruan hitz egingo da.

4.4.1.1 Aurreprozesua

Metodoaren lehen pausuak, testu lauan aurkitzen diren corpusak aurreprozesatzean datza. Testuan gertatzen diren zenbait fenomeno tratatzea beharrezkoa baita.

Prozesu hau 4.3.1.1 atalean azaltzen den TE1 sistemaren aurreprozesuaren berdina da. Bertan ere hitz anitzeko terminoak detektatu, lematizatu, eta hauen kategoriak zehazten dira.

4.4.1.2 Terminoen relebantzia neurtzen

Behin termino hautagaiak ditugula, hauen relebantzia kalkulatuko da. Hau bi pausotan kalkulatzen da. Lehenik termino bat domeinu batean agertzeko probabilitatea kalkulatzen da:

$$p(t|d_i) = \frac{freq(t \in d_i)}{\sum_{i=1..n} freq(t \in d_i)} \tag{4.6}$$

Hau egiteko, lehenik t terminoaren maiztasuna lortzen da d_i domeinu barruan. Ondoren terminoak domeinu guztietan batera duen maiztasuna lortzen da. Bi hauek zatituz lortzen da termino baten domeinu bateko probabilitatea. Probabilitate hau handia izango da, t terminoa d_i domeinuan asko agertzen bada beste domeinuekin konparatuz. n domeinu kopurua da.

Behin probabilitatea dugularik relebantzia lortu nahi da. Honen helburua d_i domeinuak t terminoak duen pisua, termino horren pisu handienarekin konparatzea da. Horrela probabilitatea handiena duen domeinuan bateko pisua emango zaio, eta beste domeinutan 1 edo txikiagoa. Hau egiteko jarraian aurkezten den formula erabiltzen da.

$$rlv(t|d_i) = \frac{p(t|d_i)}{\max_{i=1..n} p(t|d_i)} \tag{4.7}$$

Hau da, t terminoaren d_i domeinuko probabilitateari, t termino horrek izan duen probabilitate handiena zatitzen zaio.

4.4.1.3 Terminoen kontsentsua neurtzen

Hurrengo pausoa terminoak domeinu bateko dokumentuen artean nola banatuta dauden begiratzea da. Termino bat asko agertzen bada, baina agerpen guztiak dokumentu berean badaude, zarata izateko probabilitatea handia izango du. Aldiz, termino hori modu homogeneo batean banatuta badago domeinuan zehar, termino esanguratsua izateko probabilitate gehiago du. Fenomeno hau modelatzeko terminoen kontsentsua kalkulatu da.

Estatistiko hau entropian oinarritua dago, zenbat eta homogeneoago banatuta egon datuak orduan eta balio altuagoak itzultzen ditu. Bi pausotan kalkulatzen da. Lehenik domeinu barneko k dokumentu bakoitzeko terminoekin probabilitateak lortzen dira.

$$p(t|k_i) = \frac{freq(t \in k_i)}{\sum_{i=1..n} freq(t \in k_i)} \tag{4.8}$$

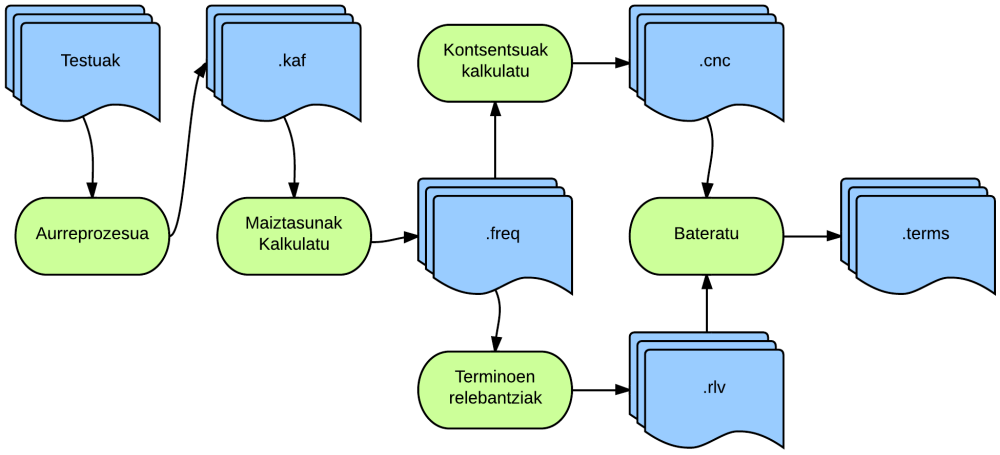
Probabilitate hauek terminoaren k dokumentuko maiztasuna domeinuko dokumentu guztietan duen maiztasunarekin zatituz lortzen dira.

Behin probabilitate guztiak kalkulatuta hauen entropia kalkulatzen da. Honekin probabilitateen homogeneotasun edo kontsentsua kalkulatu nahi da.

$$cnc(t, k_i) = H(p(t|k_j)) = \sum_{k_j \in d_i} \log_2(\frac{1}{p(t, k_j)}) \tag{4.9}$$

4.4.2 Aplikazioaren diseinua

Corpusak oso handiak izan daitezke, eta hauekin lan egiteak denbora aldetik kostu handia suposa dezake. Horregatik, terminoak erauzteko programa modu banatu batean exekuta daitezkeen azpiprogrametan banatzea erabaki da. Azpiprograma horietako bakoitzak fitxategi batzuetatik irakurtzen ditu datuak eta bere emaitza ere fitxategi ezberdinetan utziko du. Horrela prozesu bakoitza modu independente batean exekutatzea ahalbidetzen da, eta jada kalkulatuta dauden datuak berrerabiltzeko aukera ematen du. Programa nagusian honako azpi programak identifikatu dira:



Irudia 4.3: TE2en azpiprogramen egitura

Aurreprozesua(Komuna)
Prozesu hau komuna da bai TE1 eta bai TE2 termino erauzleetan. Informazio gehiagorako ikus 4.3.2.1 atala.

Maiztasunak lortu(Komuna)
Prozesu hau komuna da bai TE1 eta bai TE2 termino erauzleetan. Informazio gehiagorako ikus 4.3.2.1 atala.

Relebantziak kalkulatu
Prozesu honek corpus bakoitzaren maiztasun fitxategiak hartu eta bertan agertzen diren termino bakoitzaren relebantziak neurtzen ditu 4.4.1.2 atalean azaltzen den formula aplikatuz. Emaitza ".rlv"luzapeneko fitxategietan gordetzen da eta korpus bakoitzarentzat emaitza ezberdina gordetzen du.

Kontsentsuak kalkulatu
Prozesu honek corpus bakoitzaren maiztasun fitxategiak hartu eta bertan agertzen diren termino guztien kontsentsuak kalkulatzen ditu 4.4.1.3 atalean azaltzen den formulak aplikatuz. Emaitza ".cnc"luzapeneko fitxategietan gordetzen du.

Bateratu
Prozesu honek relebantzia zein kontsentsu fitxategiak hartu eta bateratu egiten ditu, termino fitxategiak lortzeko.

4.5 Ebaluazioa

Jarraian, sorturiko bi termino erauzleen ebaluazio bat aurkezten da. Ebaluazio honetan bi termino erauzleak konparatzeaz gain euskararako existitzen den *Erauzterm* sistemarekin konparaketa egiten da. Ebaluazioa bi domeinu ezberdinen gainean burutua izan da. Batetik historia domeinuko corpus bat erabili da, 200.000 hitz ingurukoa. Bestetik teknologia azpidomeinuko corpus bat erabili da, hau ere 200.000 hitz ingurukoa. Aipatzekoa da Erauzterm zientzia eta teknologia arlorako garatutako tresna bat dela. Corpus hauei buruzko informazio gehiago nahi izanez gero, 3.2.2 eta 3.2.3 atalak kontsulta daitezke. Ebaluazioa eskuz zein modu automatiko batean garatua izan da.

4.5.1 Ebaluazioaren metodologia

Jarraian ebaluazioaren metodologiari buruz hitz egiten da. Hau bi modutan burutua izan da. Batetik, eskuzko ebaluazio bat burutu da sistemen kalitatea neurtzeko. Bestetik, ebaluazio automatiko bat burutu da sistemen arteko antzekotasuna neurtzeko helburuarekin. Hauen metodologia jarraian azaltzen da.

4.5.1.1 Eskuzko ebaluazioa

Eskuzko ebaluazioa burutzeko erauzle eta domeinu bakoitzetik hitz bakun zein hitz anitzeko 100 pisu handienekoak hartu dira. Horrela 12 zerrenda(2

domeinu × 3 sistema × 2 hitz aniztasun) sortu dira, bakoitza 100 terminorekin. Aukeratutako termino hauek bi ebaluatzailek eskuz ebaluatu dituzte domeinukoak diren ala ez esanez. Jarraian, bi ebaluatzaileen adostasuna neurtzeko beraien arteko konkordantzia taulak zein *Kappa* neurriak kalkulatu dira. Bukatzeko, bi ebaluatzaileak ados ez zeuden terminoak baztertuz gelditu diren termino guztiekin zerrenda oso bat sortu da, *gold standard* bat sortuz.

4.5.1.2 Ebaluazio automatikoa

Ebaluazio automatikoak hiru sistemen arteko antzekotasuna neurtzea du helburu. Honetarako sistema bakoitzak sorturiko terminoak azpimultzo ezberdinetan banatu dira, bakoitzean lehendabiziko x terminoak hartuz. Ondoren multzo hauen arteko antzekotasuna neurtu da *overlap* estatistikoa erabiliz.

Ebaluazio mota hau ez dago sistemaren kalitatea neurtzera orientatua, sistemek sortzen dituzten terminoen arteko antzekotasuna neurtzera baizik. Termino erauzleak erabiltzen dituen aplikazio bat sortuko bagenu eta sistemek antzekotasun handia balute berdin izango litzateke bat ala bestea erabiltzea. Aldiz, sistemak ez badira antzekoak gure aplikazioaren emaitzetan asko eragin lezake sistema bat ala bestea erabiltzeak. Gure kasuan sistemak antzekoak ez izateak galdera oso ezberdinak sortzea suposatu dezake.

4.5.2 Ebaluatzaileen adostasuna

Termino bat domeinu batekoa den esatea ez da erraza askotan. Horregatik ebaluazioa bi ebaluatzailek burutzea erabaki zen. Ebaluatzaile bakoitzak bere aldetik burutu du ebaluazioa eta bien emaitzak konparatu dira ebaluazioaren baliozkotasuna ziurtatzeko. Oro har, bi ebaluatzaileak adostasun ona erakutsi dute.

Adostasuna neurtzeko, konkordantzia taulak eta portzentaiak erabiltzeaz gain *Cohen's Kappa* (Cohen, 1960) neurria erabili da. *Kappa* bi ebaluatzaileren arteko adostasuna neurtzeko neurri bat da. Portzentaia hutsa erabiltzea baino sendoagoa da, ausaz gertaturiko adostasuna kontuan hartzen baitu. Jarraian agertzen den moduan kalkulatzen da.

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{4.10}$$

Bertan $Pr(a)$ ikusitako adostasun maila da eta $Pr(e)$ esperotako adostasun maila. *Landis eta Koch*-ek (1977) *kappa* balioak honela sailkatzea proposatu

zuten:

Balio tartea	**Sailkapena**
<0	adostasunik ez
0-0.20	adostasun txikia
0.21-0.40	justuko adostasuna
0.41-0.60	adostasun moderatua
0.61-0.80	adostasun handia
0.81-1	adostasun ia perfektua

Taula 4.1: Cohen's Kapparen balioen sailkapena

Jarraian ebaluatzaileen adostasunak erakusten dira modu ezberdinetan antolaturik. Lehenik domeinuekiko adostasunak erakutsiko dira, ondoren sistemarekiko adostasunak eta azkenik hitz aniztasunaren araberakoak. Buruturako proba bakoitzaren datu gehiago jakin nahi izanez gero jo II eranskinera.

4.5.2.1 Adostasunak domeinuarekiko

Jarraian ebaluatzaileen arteko adostasunak domeinuka banaturik agertzen dira, lehenik historia domeinuan duten antzekotasuna erakutsiko da eta ondoren teknologia domeinukoa.

	2. Ebaluatzailea			
1. Ebaluatzailea		Bai	Ez	
	Bai	355	40	395
	Ez	75	128	203
		430	168	598

Taula 4.2: Konkordantzia taula: Historia domeinua

Historia domeinuan ebaluatzaileek izan duten *kappa* 0,55ekoa izan da. Adostasunari dagokionez %81ekoa izan da. Taulan ikus daitekeen moduan 115 hitzetan gertatu da desadostasuna eta 483 terminotan ados egon dira bi ebaluatzaileak. 600era iristeko falta diren 2 terminoak ebaluatzaile batek ezin izan zituelako etiketatu falta dira. Bi terminoen anbiguotasunaren ondorioz ezinezkoa zen domeinukoak ziren edo ez esatea.

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	122	37	159
	Ez	57	384	441
		179	421	600

Taula 4.3: Konkordantzia taula: Teknologia domeinua

Teknologia domeinuan ebaluatzaileek izan duten *kappa* 0,61ekoa izan da. Adostasunari dagokionez %84ekoa izan da. Taulan ikus daitekeen moduan 94 hitzetan gertatu da desadostasuna eta 506 terminotan ados egon dira bi ebaluatzaileak.

4.5.2.2 Adostasunak sistemarekiko

Jarraian ebaluatzaileen arteko adostasunak sistemen arabera banaturik erakusten dira. Lehenik TE1 eta TE2 sistemen adostasunak erakutsiko dira eta azkenik Erauzterm sistemarena.

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	128	47	175
	Ez	47	175	222
		175	222	400

Taula 4.4: Konkordantzia taula: TE1

TE1 sisteman ebaluatzaileek izan duten *kappa* 0,51ekoa izan da. Adostasunari dagokionez %76ekoa izan da. Taulan ikus daitekeen moduan 94 hitzetan gertatu da desadostasuna eta 306 terminotan ados egon dira bi ebaluatzaileak.

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	168	14	182
	Ez	33	185	218
		201	199	400

Taula 4.5: Konkordantzia taula: TE2

TE2 sisteman ebaluatzaileek izan duten *kappa* 0,77ekoa izan da. Adostasunari dagokionez %88ekoa izan da. Taulan ikus daitekeen moduan 47 hitzetan gertatu da desadostasuna eta 353 terminotan ados egon dira bi ebaluatzaileak.

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	180	16	196
	Ez	52	152	204
		232	168	400

Taula 4.6: Konkordantzia taula: Erauzterm

Erauzterm sisteman ebaluatzaileek izan duten *kappa* 0,66ekoa izan da. Adostasunari dagokionez %83ekoa izan da. Taulan ikus daitekeen moduan 68 hitzetan gertatu da desadostasuna eta 332 terminotan ados egon dira bi ebaluatzaileak.

4.5.2.3 Adostasunak hitz aniztasunarekiko

Jarraian ebaluatzaileen adostasuna hitz aniztasunaren arabera antolatuta erakusten da. Lehenik hitz bakarreko terminoak kontuan harturik ebaluatzaileek erakutsi duten adostasunaren datuak erakusten dira. Ondoren berdina egiten da hitz anitzeko terminoekin.

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	204	41	245
	Ez	80	275	355
		284	316	600

Taula 4.7: Konkordantzia taula: Hitz bakarrekoak

Hitz bakarreko terminoak kontuan hartuta ebaluatzaileek izan duten *kappa* 0,59ekoa izan da. Adostasunari dagokionez %80ekoa izan da. Taulan ikus daitekeen moduan 121 hitzetan gertatu da desadostasuna eta 479 terminotan ados egon dira bi ebaluatzaileak.

	2. Ebaluatzailea			
1. Ebaluatzailea		Bai	Ez	
	Bai	273	36	309
	Ez	52	237	289
		325	273	598

Taula 4.8: Konkordantzia taula: Hitz anitzekoak

Hitz anitzeko terminoak kontuan hartuta ebaluatzaileek izan duten *kappa* 0,70ekoa izan da. Adostasunari dagokionez %85ekoa izan da. Taulan ikus daitekeen moduan 88 hitzetan gertatu da desadostasuna eta 510 terminotan ados egon dira bi ebaluatzaileak.

4.5.2.4 Adostasunak guztira

	2. Ebaluatzailea			
1. Ebaluatzailea		Bai	Ez	
	Bai	477	77	554
	Ez	132	512	644
		609	589	1198

Taula 4.9: Konkordantzia taula: Historia domeinua

Proba guztiak kontuan hartuta ebaluatzaileek izan duten *kappa* 0,65ekoa izan da. Adostasunari dagokionez %82ekoa izan da. Taulan ikus daitekeen moduan 209 hitzetan gertatu da desadostasuna eta 989 terminotan ados egon dira bi ebaluatzaileak.

4.5.2.5 Ondorioak

Datuak ikusirik, *kappa* zein antzekotasun neurriek ebaluazioa baliozkoa izan dela ondorioztatzen dute. *Kappa* beti mantendu da 0,5 eta 0,8 artean eta hau neurri ona da modu honetako sistema baten ebaluaziorako.

4.5.3 Gold Standard-ak

Bi ebaluatzaileen emaitzetatik bi *gold standard* sortu dira. Hauek bi ebaluatzaileak ados zeuden terminoekin osatu dira. Bai biek domeinuko terminotzat jo dituzten terminoak eta bai biek domeinuz kanpoko terminotzat jo

dituzten terminoak erabili dira hauek eratzeko. Jarraian erakusten dira *gold standard*aren zenbait datu.

4.5.3.1 Historia domeinua

	Termino Bakunak	Hitz Anitzeko Terminoak	Guztira
Egokiak	129	181	310
Okerrak	60	66	126
Guztira	189	247	436

Taula 4.10: Historia domeinuaren GoldStandard-aren estatistikak

Guztira 436 termino pasa dira historia domeinuko *gold standard*-a eratzera. Hauetatik 310 domeinukoak dira eta 126 domeinuz kanpokoak. Bestalde, 189 termino bakunak dira eta 247 hitz anitzekoak.

4.5.3.2 Teknologia domeinua

	Termino Bakunak	Hitz Anitzeko Terminoak	Guztira
Egokiak	40	79	119
Okerrak	193	164	357
Guztira	233	243	476

Taula 4.11: Historia domeinuaren GoldStandard-aren estatistikak

Guztira 476 termino pasa dira teknologia domeinuko *gold standard*-a eratzera. Hauetatik 119 domeinukoak dira eta 357 domeinuz kanpokoak. Bestalde, 233 termino bakunak dira eta 243 hitz anitzekoak.

4.5.4 Sistemen doitasun eta estaldura

Behin *gold standard* bat eratu delarik, sistema bakoitzaren ebaluazio bat burutu da honekiko. Horrela, sistemen doitasun, estaldura eta F_1 neurria kalkulatu dira.

Doitasuna eta estaldura sistema bitar bat (bai/ez) ebaluatzeko estatistikoak dira. Doitasunak sistemak emandako emaitzetatik zenbat zuzenak

diren adierazten du. Estaldurak, aldiz, sistemak egoki eman zitzakeen emaitza guztietatik zenbat eman dituen adierazten du. Gure kasuan, doitasunak sistemak emandako terminoetatik zenbat diren zuzenak adierazten du. Estaldurak *gold standard*ean zeuden termino egokietatik sistemak zenbat eman dituen emaitzatzat adierazten du. Honakoak lirateke ebaluazioan erabilitako doitasun eta estaldura formulak.

$$Doitasuna = \frac{|sistemak\ emandako\ emaitza\ egokiak\ |}{|sistemak\ emandako\ emaitzak\ |} \tag{4.11}$$

$$Estaldura = \frac{|sistemak\ emandako\ emaitza\ egokiak\ |}{|gold\ standardeko\ emaitza\ egokiak\ |} \tag{4.12}$$

Ikus daitekeen moduan, estaldura *gold standard*arekiko kalkulatzen da, ezin baita jakin benetan zenbat termino dauden domeinuan. Honek beraz, ez digu estaldura erreala emango, baina bai honen nozio bat.

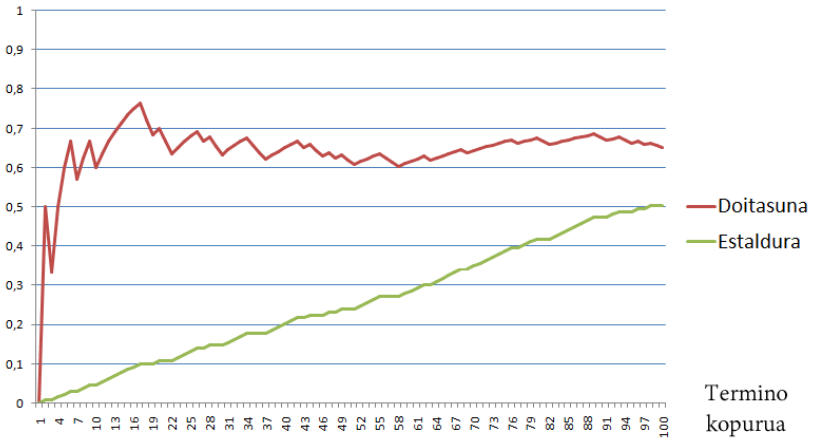
F_1 neurria doitasun eta estalduraren arteko konbinazio bat da. Honela kalkulatzen da:

$$F_1 = 2 \cdot \frac{estaldura \cdot doitasuna}{estaldura + doitasuna} \tag{4.13}$$

Ebaluazio hau sistema, domeinu eta hitz aniztasunarekiko burutua izan da. Jarraian kategoria bakoitzean ondoen ibili diren sistemen datuak aurkeztuko dira. Datu gehiago ikusi nahi izanez gero, II eranskina kontsulta daiteke.

4.5.4.1 Historia (hitz bakarrekoak)

Historia domeinuan, hitz bakarreko terminoei dagokionez Erauzterm sistema izan da emaitza onenak lortu dituena.

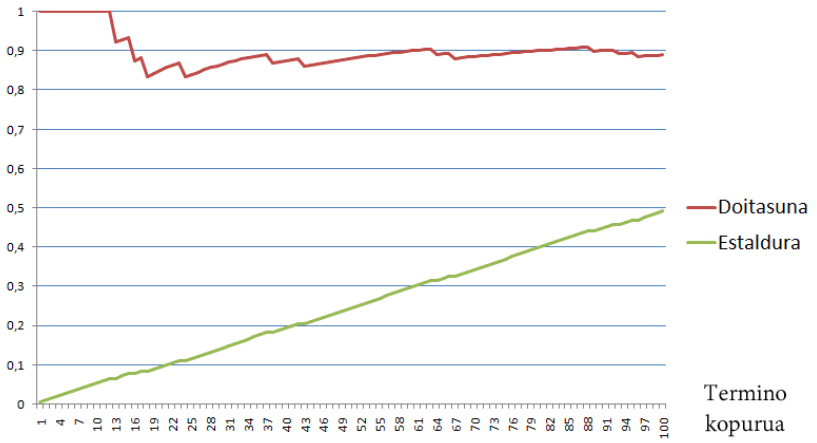


Irudia 4.4: Ebaluazio grafikoa: Erauzterm, historia, hitz bakarrekoak

Grafikoan terminoak 1-100 doazen heinean doitasun eta estaldura balioa nola mugitzen diren ikus daiteke. 100 termino harturik, bere doitasuna 0,65 da eta bere estaldura 0,50. F_1 neurriari dagokionez bere balioa 0,57 da. Estaldura eta doitasuna grafikotik kanpo mozten dira, honek sistemak termino ugari doitasun onez erauzten dituen ziurtasuna ematen digu. Bestetik doitasunaren joera mantentzekoa dela dirudi, hau ona izango da termino asko behar dituzten sistemetan erabili ahal izateko.

4.5.4.2 Historia (hitz anitzekoak)

Historia domeinuan, hitz anitzeko terminoei dagokionez TE2 sistema izan da emaitza onenak lortu dituena.

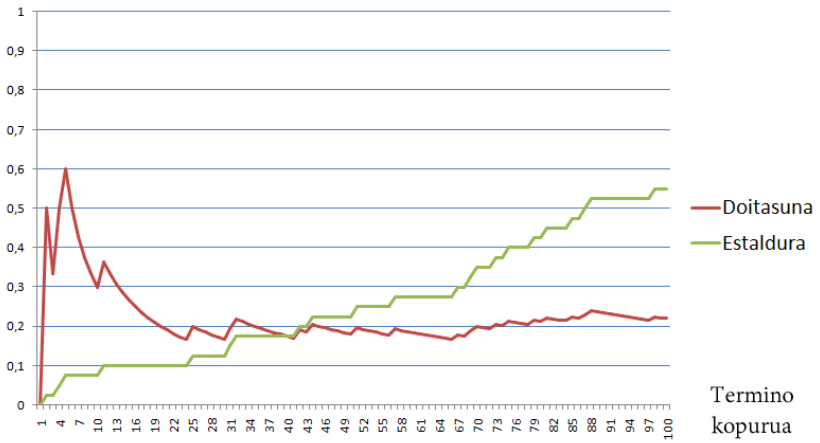


Irudia 4.5: Ebaluazio grafikoa: TE2, historia, hitz anitzekoak

Grafikoan terminoak 1-100 doazen heinean doitasun eta estaldura balioa nola mugitzen diren ikus daiteke. 100 termino harturik, bere doitasuna 0,89 da eta bere estaldura 0,49. F_1 neurriari dagokionez bere balioa 0,63 da. Doitasun eta estaldura ez dira grafikoaren barnean ebakitzen, honek sistemaren kalitatea mailaren ikuspegi bat ematen digu. Estaldura gelditu gabe gora doan bezala, doitasuna nahiko finko mantentzen da 0.9 inguruan. Honek termino gehiago hartuta ere horrela mantenduko diren irudipena ematen du.

4.5.4.3 Teknologia (hitz bakarrekoak)

Teknologia domeinuan, hitz bakarreko terminoei dagokionez TE1 sistema izan da emaitza onenak lortu dituena.

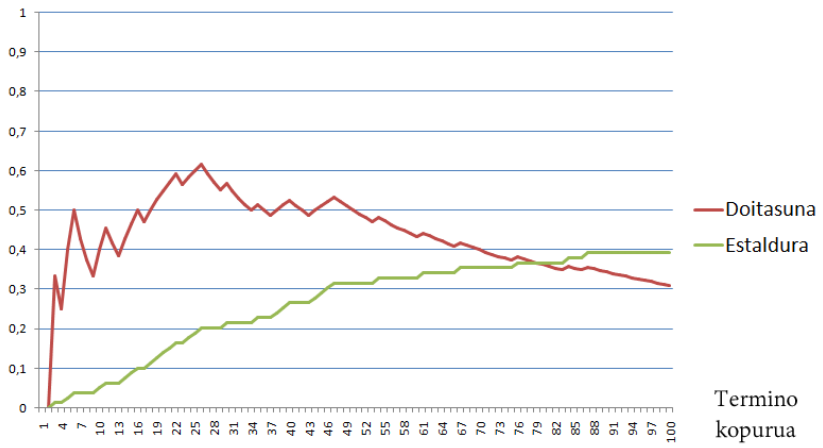


Irudia 4.6: Ebaluazio grafikoa: TE1, teknologia, hitz bakarrekoak

Grafikoan terminoak 1-100 doazen heinean doitasun eta estaldura balioa nola mugitzen diren ikus daiteke. 100 termino harturik, bere doitasuna 0,22 da eta bere estaldura 0,55. F_1 neurriari dagokionez bere balioa 0,31 da. Estaldura eta doitasunaren ebakidura puntua nahiko goiz gertatzen da. Honek sistemaren doitasuna txikia dela irudikatzera ematen gaitu, estaldura igotzen hasi ahala bizkor jaisten baita hau.

4.5.4.4 Teknologia (hitz anitzekoak)

Teknologia domeinuan, hitz anitzeko terminoei dagokionez TE1 sistema izan da emaitza onenak lortu dituena.



Irudia 4.7: Ebaluazio grafikoa: TE1, teknologia, hitz anitzekoak

Grafikoan terminoak 1-100 doazen heinean doitasun eta estaldura balioa nola mugitzen diren ikus daiteke. 100 termino harturik, bere doitasuna 0,31 da eta bere estaldura 0,39. F_1 neurriari dagokionez bere balioa 0,34 da. Doitasuna eta estaldura erlatiboki bizkor ebakitzen dira historia domeinuarekin konparatuz. Honek sistemaren doitasuna azkar jaisten dela ondorioztatzera eramaten gaitu. Gainera honek behera jaisten jarraituko duen itxura du 100 terminotatik aurrera jarraituz.

4.5.4.5 Ondorioak

Emaitzak ikusirik, garatutako TE1 eta TE2 sistemen emaitzak onak direla esan daiteke. Izan ere, probak buruturiko 4 kategoriatarik (hitz anitzeko eta bakarrekoak historia eta teknologia domeinuan), 3tan hauen emaitzak izan dira hoberenak. Hala ere azpimarragarria da, doitasunak teknologia domeinuan izan duen beherakada. Historia domeinuan 0,65 eta 0,89 izatetik, teknologia domeinuan 0,22 eta 0,31 izatera pasa da. Honek emaitzak hurbilagotik begiratzean bultzatu gaitu eta corpus honetan dokumentu pare baten aldetik zarata handia sortu izan daitekeela suposatzen dugu. Bertan hizkuntza zerrenda batzuk aurkitu dira, hizkuntzaren prozesamenduari buruz hitz egiten den dokumentu bat baino gehiagotan errepikatzen direnak. Beraz teknologia domeinuaren gainean sorturiko datuak tentuz hartu beharrekoak direla uste dugu.

Bestalde, azpimarragarria da baita ere, TE1 sistemaren jokaera zarataren aurrean. WordNet zerbitzua erabili izanak hitz zaratatsu asko baztertzera eraman du, eta hortik bere emaitza onak teknologia domeinuan.

Hitz bakun eta anitzak konparatuz gero hitz anitzeko terminoetan doitasuna hobea dela ikus daiteke. Ebaluatzaileek emandako iritzi pertsonalean,

hitz anitzeko terminoak etiketatzeko errazagoak zirela esan zuten. Irudipen hau doitasunarekin emaitzekin berretsi da.

Laburbilduz, TE1 eta TE2 sistemak maila onean mantendu dira Erauzterm sistemaren ondoan eta TE1 sistemak corpus zaratatsuen aurrean izan dezakeen sendotasuna frogatu du. Wordnet moduko hiztegi bat erabiltzeak zenbait termino zaratatsu baztertzen lagundu dio sistema honi.

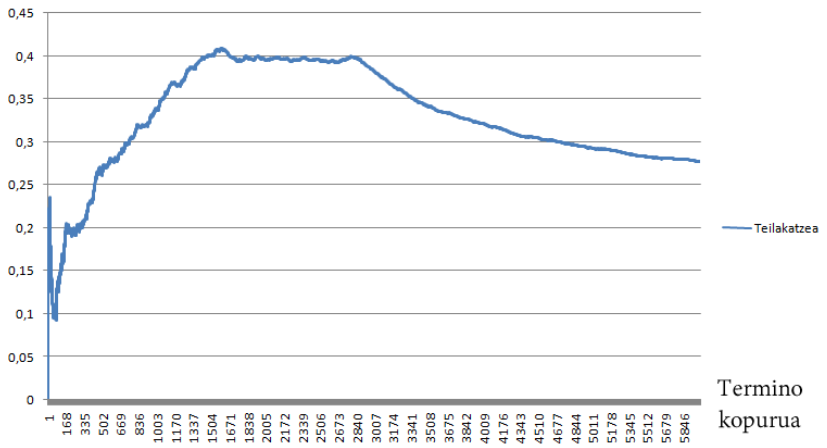
4.5.5 Sistemen antzekotasuna

Jarraian, sistemek sortzen dituzten terminoen arteko antzekotasuna neurtu da. Bertan helburua ez da sistemen kalitatea neurtzea, hauek sortzen dituzten terminoak zein mailatan berdinak diren neurtzea baizik. Horretarako *overlap* neurria erabili da. *Overlap* neurriak bi multzo zein neurritan teilakatzen diren adierazten du. X eta Y multzoak izanik, honela kalkulatuko genuke $overlap(X, Y)$:

$$overlap(X, Y) = \frac{|X \cap Y|}{min(|X|, |Y|)} \tag{4.14}$$

4.5.5.1 Historia domeinua

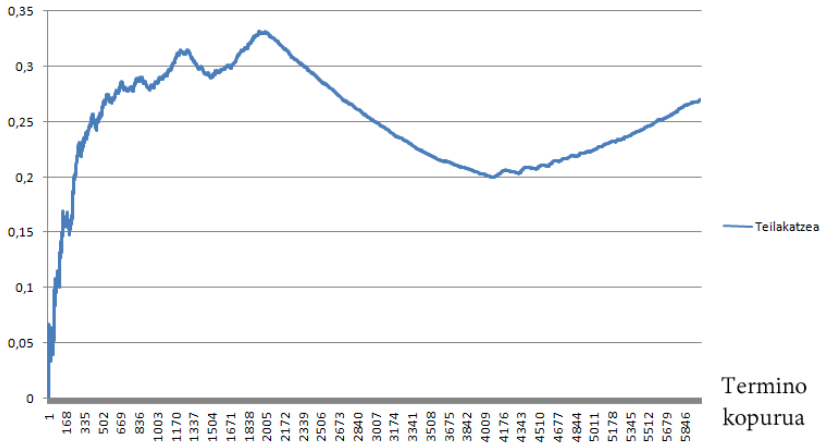
Jarraian historia domeinuan buruturiko probetan hiru sistemek izan dituzten antzekotasunak erakusten dira.



Irudia 4.8: Teilakatze grafikoa: TE1 eta TE2, historia domeinua

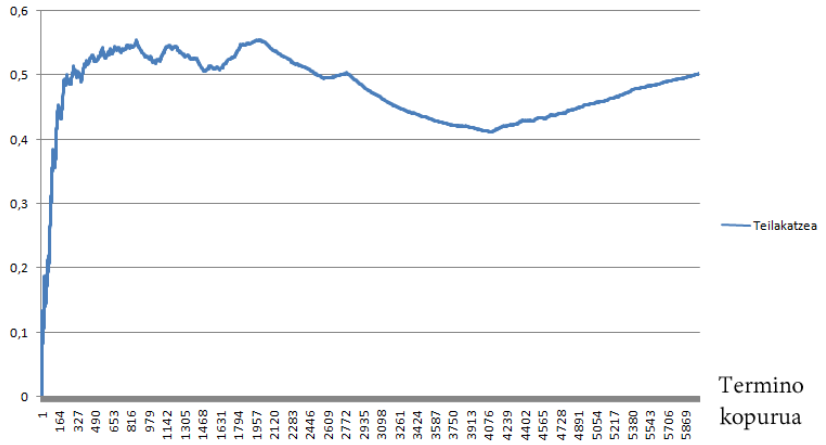
Goiko grafikoan TE1 eta TE2 sistemek historia domeinuan erauzitako terminoek duten antzekotasuna ikus daiteke. Erauziriko termino guztiak

konparatuz antzekotasuna 0,27an kokatzen da. Honek bien arteko antzekotasuna ez dela oso handia erakusten du.



Irudia 4.9: Teilakatze grafikoa: TE1 eta Erauzterm, historia domeinua

Goiko grafikoan TE1 eta Erauzterm sistemek historia domeinuan erauzitako terminoek duten antzekotasuna ikus daiteke. Erauziriko termino guztiak konparatuz antzekotasuna 0,26an kokatzen da. Honek bien arteko antzekotasuna ez dela oso handia erakusten du.

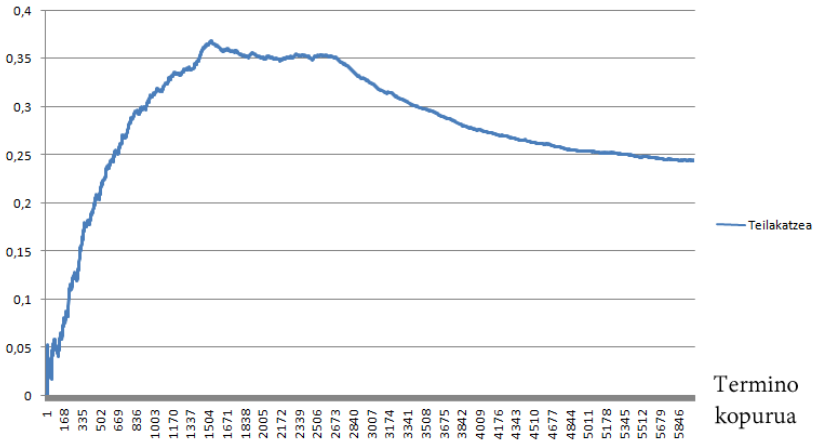


Irudia 4.10: Teilakatze grafikoa: TE2 eta Erauzterm, historia domeinua

Goiko grafikoan TE2 eta Erauzterm sistemek historia domeinuan erauzitako terminoek duten antzekotasuna ikus daiteke. Erauziriko termino guztiak konparatuz antzekotasuna 0,50ean kokatzen da. Bi sistemen arteko antzekotasuna oso handia ez den arren kontuan hartzekoa da.

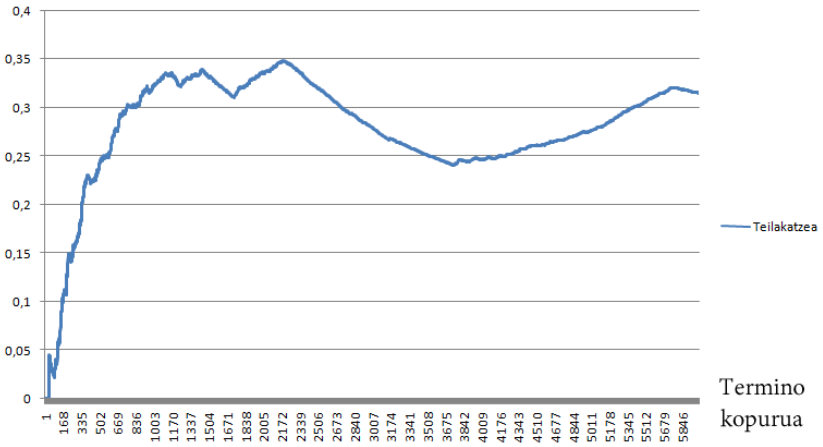
4.5.5.2 Teknologia domeinua

Jarraian teknologia domeinuan buruturiko probetan hiru sistemek izan dituzten antzekotasunak erakusten dira.



Irudia 4.11: Teilakatze grafikoa: TE1 eta TE2, teknologia domeinua

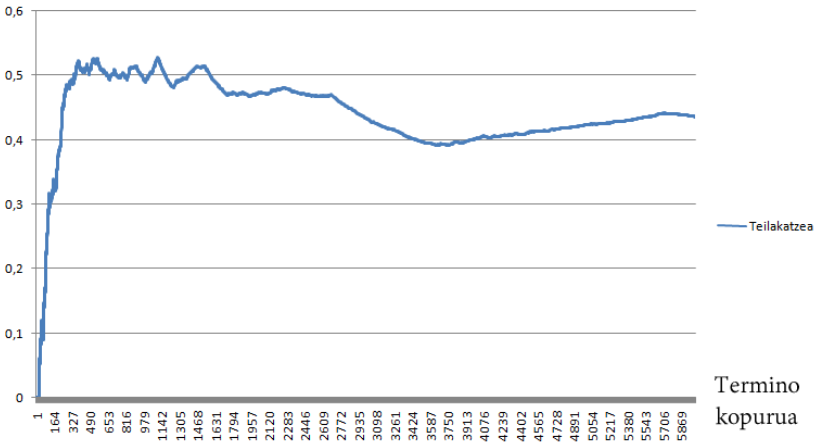
Goiko grafikoan TE1 eta TE2 sistemek teknologia domeinuan erauzitako terminoek duten antzekotasuna ikus daiteke. Erauziriko termino guztiak konparatuz antzekotasuna 0,24an kokatzen da. Honek bien arteko antzekotasuna ez dela oso handia erakusten du.



Irudia 4.12: Teilakatze grafikoa: TE1 eta Erauzterm, teknologia domeinua

Goiko grafikoan TE1 eta Erauzterm sistemek teknologia domeinuan erauzitako terminoek duten antzekotasuna ikus daiteke. Erauziriko termino guz-

tiak konparatuz antzekotasuna 0,32an kokatzen da. Honek bien arteko antzekotasuna ez dela oso handia erakusten du.



Irudia 4.13: Teilakatze grafikoa: TE2 eta Erauzterm, teknologia domeinua

Goiko grafikoan TE2 eta Erauzterm sistemek teknologia domeinuan erauzitako terminoek duten antzekotasuna ikus daiteke. Erauziriko termino guztiak konparatuz antzekotasuna 0,43an kokatzen da. Bi sistemen arteko antzekotasuna oso handia ez den arren kontuan hartzekoa da.

4.5.5.3 Ondorioak

Datuak ikusirik, sistemak orokorrean 0,3ko antzekotasun bat aurkezten dute, beraz nahiko emaitza ezberdinak ematen dituzte. Hala ere nabarmentzekoa da TE2 eta Erauzterm-en arteko antzekotasuna(0.5 inguru). Bai historia, bai teknologia domeinuan detektatu dituzten terminoen erdia konpartitzen dute gutxi gora behera. Antzekotasun hau erabiltzen duten algoritmoaren arrazoiz gertatu dela suposatzen dugu. Biek *tf.idf* estiloko estatistiko bat erabiltzen baitute. TE1 sistemak ere horrelako estatistiko bat erabiltzen duen arren, *WordNet* zerbitzua erabili izanak besteekiko ezberdindu du.

4.5.6 Ondorio orokorrak

Orokorrean ebaluatzaileek adostasun maila egokia erakutsi dute. Honek frogak baliozkoak izan direla ondorioztatzera eramaten gaitu.

Sistemen kalitateari dagokionez garaturiko bi sistemak existitzen zen Erauzterm sistemaren parean eta kasu batzuetan hobeto dabiltzala ikusi da.

Teknologia eta historia domeinuen gainean sistemek izan duten portaera oso ezberdina izan da, hiru sistemak historia domeinuan emaitza hobeak eman dituztelarik. Honek teknologia domeinuan beste ebaluazio bat egin beharra aurreikusten du etorkizunerako, bertan agertu den zarata arazoa konpondu ostean.

Hitz bakar eta hitz anitzeko terminoak konparatzen baditugu bigarrenekin emaitza hobeak lortu direla ikus daiteke. Froga garatu zenean ebaluatzaileen iritzia hitz anitzeko terminoak domeinukoak ziren edo ez esateko errazagoak zirela izan zen. Iritzi hau hitz anitzeko terminoen emaitza onekin berretsi da.

Azkenik, termino hutsak erauzi beharrean adieraz erauzteak izan ditzakeen onurak nabarian gelditu dira. Zarataren aurrean adierak erauzteko sistemak erakutsi duen sendotasuna nabarmentzekoa da.

5 Kapitulua

Galdera sortzailea

Gaien Aurkibidea

Galdera sortzailea proiektuan jorratu diren bi ekarpen nagusietako bat da. Bertan galdera motaren aukeraketa eta galdera eraikuntza burutzen dituen modulua aurkezten da. Lehenik galdera sortzaile automatikoen inguruan burutu den aurrekarien azterketa azalduko da. Ondoren galdera sortzaile moduluaren garapenari buruz hitz egingo da. Amaitzeko, burutu den ebaluazioaren inguruko azalpenak emango dira.

5.1 Analisia

Jarraian galdera sortzaile automatikoaren inguruko analisia aurkezten da. Bertan aurrekarien azterketa, aplikazioaren egitura eta bere fase guztiak azalduko dira.

5.1.1 Aurrekarien azterketa

Jarraian galdera sortzaile automatikoen inguruan burutu den aurrekarien azterketari buruz hitz egiten da. Lehenik beste hizkuntzetarako aurrekari azterketa bat azalduko da, eta ondoren, euskararako dauden sistemen inguruan hitz egingo da.

5.1.1.1 Beste hizkuntzetarako aurrekari azterketa

Galdera sorkuntza automatikoak irakaskuntzan izan ditzakeen aplikazioak direla medio, ikerketa talde askotan landu da gai hau. Azterketa bibliografikoa burutu ondoren, honako sistemak izan dira interesgarritzat jo ditugunak.

Patroiak ikasteko teknikak
Curto S. et al.-ek (2012) galdera sorkuntzarako transformazio erregelak automatikoki sortzen dituen sistema bat proposatzen dute. Transformazio erregela bat, esaldi bati aplikatzen zaion erregela bat da, hau galdera bihurtzeko helburuz. Erregela hauek ez dute galderak sortzeko soilik balio, erantzunak zein distraktoreak ere sortzeko gai dira. Beraien sistema bi fasetan banatzen dute: patroi ikasketa eta galdera sorkuntza.

Patroi ikasketa
Sistemak galdera/erantzun pareak jasotzen ditu entrenamendu gisa. Galdera hauek chunk mailan sintaktikoki analizatzen dira. Horrela galdera zein erantzuna osatzen duten chunkak lortzen dira. Adibidez:

Galdetzailea	aditz sintagma	izen sintagma	?	erantzuna
Who	composed	Moonlight Sonata	?	Beethoven

Taula 5.1: Galdera analisi sintaktikoa burutu ondoren

Behin esaldia chunketan banatuta dagoela, galderari galdetzailea kentzen zaio. Galderan gelditu dena eta erantzuna erabiliz permutazio posibleak sortzen dira. Permutazioetan "*"ikurrak ere gehitzen dira, hauen ordez edozer ager daitekeela adieraziz. Hona hemen aurreko esaldiaren permutazio adibide batzuk:

1	composed	Moonlight Sonata	*	Beethoven	
2	**Moonlight Sonata**	*****	**composed**	*****	**Beethoven**
3	Beethoven	Moonlight Sonata	*	composed	
4	**Beethoven**	**composed**	**Moonlight Sonata**		
5	...	...	...	...	...

Taula 5.2: Permutazio posibleak

Behin permutazio guztiak lortuta, hauek bilatzaile baten bidez kontsultatzen dira sarean. Emaitza kopuru minimo bat baino gehiago aurkitzen bada permutazioa gorde egiten da, bestela albo batera uzten dira. Goiko permutazioetatik 2. eta 4. dira adibidez permutazio ohikoenak. "Moonlight Sonata was composed by Beethoven"eta "Beethoven composed Moonlight Sonata"esaldiak existitzen baitira.

Behin zein patroi diren posible identifikatu ondoren, hauek galdera sorkuntzarako gordetzen dira. Honetarako ez dira hitzak bere horretan gordetzen, berauen informazio sintaktikoa baizik. Helburua ez baita behin eta berriz Beethoven-en inguruko galderak sortzea, egitura sintaktiko hori duten galderak sortzea baizik. Honakoak lirateke goiko adibidetik gordetako patroiak:

2.patroia(esaldia)	izen sintagma	*	aditz sintagma	*	Erantzuna
2.patroia(galdera)	galdetzailea	aditz sintagma	izen sintagma		
4.patroia(esaldia)	erantzuna	aditz sintagma	izen sintagma		
4.patroia(galdera)	galdetzailea	aditz sintagma	izen sintagma		

Taula 5.3: Gordetako patroiak

Galdera sorkuntza
Behin patroiak ikasita sistema prest dago galderak sortzen hasteko. Galderak sortzeko jaso den testua sintaktikoki analizatzen da, galdera/erantzun pareekin egiten zen bezala. Behin testua chunketan banaturik dagoela patroiak testuan bilatzen dira. Parekatutako patroi guztiak transformatzen dira bere erregelaren bitartez eta galderak sortzen dira.

Sortutako galderetan emaitzak hobetzeko asmoz filtro batzuk ere inplementatu dituzte. Hala nola, anafora filtro bat, hau da, esaldian kanpo erreferentziarik (hori, hau, aurretik aipaturiko...) badago baztertu egiten da. Filtro semantiko bat ere inplementatu dute. Wordnet erabiliz erantzunaren kategoriak detektatzen dituzte (lekua, pertsona...). Hauek ez badatoz bat galdetzailearekin galdera albo batera uzten dute.

Galderen kalitatea hobetzeko teknikak
Heilman M. et al.-ek (2010) proposaturiko sistemaren helburu nagusiena galdera ahalik eta gramatikalki zuzenena lortzean datza. Honetarako, lehenik galderak erregeletan oinarrituriko sistema batekin sortzen dira. Ondoren *machine learning* teknikez baliatuz galderak sailkatzen dira, hauetatik hoberenak aurkitzeko asmoz.

Lehen pausoa testua sintaktikoki analizatzea da. Ondoren testuan agertzen diren esaldiak sinplifikatzen[1] dira. Behin esaldiak sinplifikatuta esaldiak galderetan transformatzen dira. Honetarako tregex eta tsurgeon patroiak erabiltzen dira.

Tregex patroiak sintaktikoki analizaturiko esaldi batean bilaketak burutzeko balio dute. XML dokumentu batean Xpath teknologiak lagunduko ligukeen modu antzekoan. Horrela esaldiaren zuhaitzean galdera sortzeko beharrezkoak diren nodoak bilatu daitezke.

Tsurgeon erregelak tregex patroiekin konbinaketan lan egiten dute. Beren helburua tregex patroiek bilaturiko nodoen gainean transformazioak burutzea da, XML dokumentuetan XSLT transformazio lengoaiak egingo lukeen modu antzeko batean.

Bi hauek konbinatuz eskuzko erregela ezberdinak idatzi dira eta honek

[1]Sinplifikazioa hizkuntzaren prozesamenduko arlo bat da. Bere helburua esaldiak hartu eta esanahi berdina mantenduz modu sinpleagoan jartzea da. Adibidez, mendeko perpausak dituzten esaldiak bitan banatuz.

esaldi bat automatikoki galdera bihurtzeko balio du.

Behin galderak sortuta dituztelarik hauek sailkatzea da egin beharreko bakarra. Honetarako sailkatzaile estatistikoa eskuz etiketatuko corpus baten gainean entrenatu dute. Galderen kalitatea neurtzeko hurrengo ezaugarriak erabili dituzte batzuk aipatzearren:

- Esaldiaren luzera
- Erabiltzen den galdetzailea
- Ezezko hitzik agertzen den esaldian("inoiz", "ez"...)
- Informazio gramatikala: zenbat aditz, zenbat adjektibo...
- Zein transformazio erabili den galdera sortzeko

Guztira 53 ezaugarri ezberdin erabili dituzte sailkatzailearentzat. Behin hau entrenatuta galdera bakoitza sailkatzen da eta bere pisuaren arabera ordenatzen da.

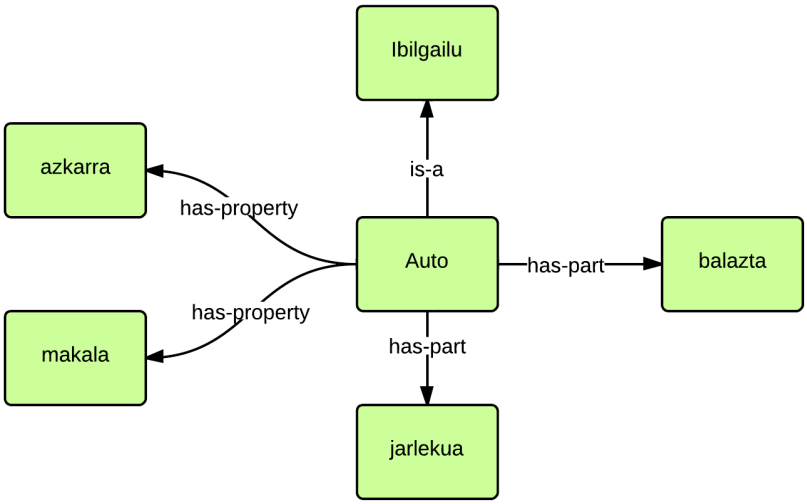
Sistemak sortzen dituen galderak sailkatzailea erabili gabe %27an dira onargarriak. Aldiz, sailkatzaileak emandako pisuaren arabera ordenatu eta hauetatik %20 pisu handieneko galderak hartuz gero, hauen onargarritasun maila %52koa da.

Galdera sortzaileen ebaluazio automatikoa
Bernhard D. et al.-ek (2012) aurreko sistemaren oso antzekoa den sistema bat proposatzen dute. Analisi sintaktikoa, sinplifikazioa eta eskuzko erregelak. Honen berezitasuna ez da sisteman aurkitzen, berau ebaluatzeko moduan baizik. Galdera sortzaileak modu automatikoan ebaluatzeko metodo bat proposatzen dute. Honetarako galderak automatikoki erantzuteko sistementzako datuak erabiltzen dituzte. CLEF lehiaketako entrenamendu corpusak erabiltzen dituzte honetarako. Corpus hauetan testu batzuen inguruko galdera eta erantzunak agertzen dira. Ebaluazioan testu hauek hartu eta galderak sortzen dituzte. Ondoren, entrenamendu corpusean zeuden galdera/erantzunak eta sistemak sorturiko galdera/erantzunak konparatzen dituzte.

Kontzeptu mapetan oinarrituriko galdera sorkuntza
Andreq M et al.-ek (2012) proposaturiko sisteman galderak esaldi mailan sortu beharrean testu mailan sortzen dituzte. Honetarako testuan dagoen

informazioaren mapa kontzeptual bat eratzen dute eta ondoren mapa honetatik sortzen dituzte galderak.

Mapa kontzeptuala sortzeko 21 erlazio ezberdin erabiltzen dituzte, horien artean: "is-a", "has-property", "has-consequence","reason","implies","outcome"eta "means". Jarraian erakusten den moduko mapa kontzeptualak sortzen dituzte:



Irudia 5.1: Mapa kontzeptualen adibide bat

Bertan agertzen den informazioak honakoa dio: Auto bat azkarra edo makala izan daiteke. Jarlekua eta balazta auto baten parte dira eta auto bat ibilgailu mota bat da.

Mapa kontzeptual hauek sortzeko patroi batzuk erabiltzen dituzte, adibidez "have"aditza agertzen denean badakite hortik "has-part"erlazio bat sor daitekeela eta "is"aditzarekin "has-property"edo "is-a"erlazio bat.

Behin mapa kontzeptuala sortuta galderak hemendik sor daitezke. Hau egiteko eskuz idatziriko erregelak erabiltzen dituzte.

5.1.1.2 Euskararako aurrekarien azterketa

Euskararako arlo honetako aurrekariak ez dira beste zenbait hizkuntzatarako bezain ugariak. Bertan bi aurrekari aztertu dira, bata entitate numerikoetan espezializatua eta bestea helburu orokorreko galderak sortzen dituena.

Entitate numerikoetan oinarrituriko galdera sorkuntza
Arikiturri[10] irakaskuntzan laguntzeko diseinaturiko tresna bat da. Honek, ikasleentzat modu askotariko ariketak prestatzen ditu, hutsuneak betetzeko ariketak, erroreak zuzentzeko ariketak, erantzun anitzeko ariketak... Prestatzen dituen ariketen artean ariketa mota batek testu batekiko galderak erantzutean datza.

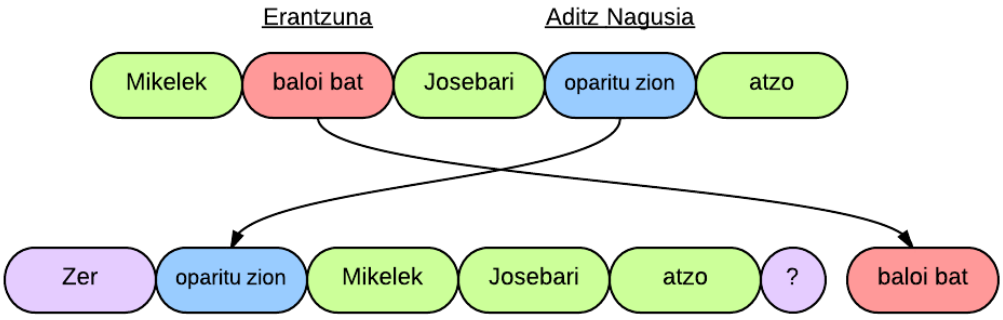
Honen ondorioz euskararako galdera sorkuntzan lehen saiakera bat burutu zuten. Saiakera hau entitate numerikoetan oinarritutakoa izan zen. Entitate numeriko bat zenbaki, sinbolo eta hitzez osaturiko multzo bat da. Adibide bezala, *%10*, 9.99 €, *ehuneko hamar* eta *bederatzi euro eta laurogeita bederatzi zentimo* entitate numerikoak dira.

Hauek detektatzeko lehenik testua maila sintaktiko zein morfologikoan analizatzen dute. Honetarako Ixati analizatzailea erabiltzen dute. Ondoren *NuERCB (Numerical Entity Recogniser and Classifier for Basque)* erremintaren bitartez testuan entitate numerikoak detektatzen dituzte. Tresna hau egoera finituko tekniketan oinarritzen da eta entitateak detektatzeaz gain bere inguruko informazioa ere ematen du, hala nola, entitateek datak, orduak edo unitateak errepresentatzen dituzten.

Behin datu guzti hauek edukirik, galderak eraikitzen dituzte. Honetarako bai NuERCB tresnak zein analizatzaileak emandako kasuaren eta chunken informazioa erabiltzen dituzte. Horrela *Zenbat kilometro...?*, *Noiz..?*, *Zein ordutan...?* moduko galderak sortzeko gai dira erregela sistema batean oinarrituta.

Seneko
Seneko[21][16] testuak landuz ikasteko aukera ematen duen aplikazio bat da. Bere berezitasuna testu bat emanda galderak automatikoki sortzeko gai izatea da. Honetarako, QG-Ixati euskararako galdera sortzaile automatikoa erabiltzen du.[9]

Galderak sortzen hasteko lehenik testuak maila sintaktiko zein morfologikoan analizatzen dira. Honetarako Ixa ikerketa taldeko chunketan oinarrituriko *Ixati* analizatzailea erabiltzen da. Behin esaldia chunketan zatituta dagoela, galderaren erantzuna izango den nodoa aukeratzen da. Honetarako izen sintagmaren kasu informazioa erabiltzen da. Kasu zehatz batzuk dituzten chunkak aukeratzen dira erantzun hautagai bezala. Kasu bereko bi chunk badaude entitate izendatua izatea hobesten dute.



Irudia 5.2: Chunketan oinarrituriko galdera eraikuntza

Behin erantzuna izango den chunka aukeratuta honen inguruan galdera sortzeari ekiten zaio. Aditz nagusia duen chunka esaldi hasierara mugitzen da eta erantzun chunka kentzen da. Honekin galderaren zatirik handiena sortuta dago.

Galdera bukatzeko hurrengo pausoa galdetzailea zein izango den erabakitzea da. Honetarako, chunkaren kasua erabiltzen da lehenik. Horrela adibidez absolutibo kasua duen erantzun batentzat galdetzailea *nor* edo *zer* izango dela ondorioztatzen dute.

Hurrengo pausoa bizidun/bizigabe hiztegiaren erabilera da. Hiztegi honetan erantzuna biziduna den edo ez jakin daiteke, eta beraz goiko kasuan *zer* edo *nor* galdetzailearen artean aukera daiteke.

Kasu askotan galdetzaile aukeraketa hor amaitzen da. Baina kasu batzuetan oraindik arazoak daude galdetzaileak zehazteko, bai bizidun/bizigabe hiztegian agertzen ez delako, bai agertuta ere oraindik bi edo aukera gehiago daudelako. Adibidez inesibo kasuan galdetzaileak, *non*, *norengan*, *noiz* izan daitezke. Erantzuna bizigabe dela ondorioztatuta ere, oraindik *non* eta *noiz* artean ezin da erabaki. Kasu horietarako rol semantikoen hiztegi bat erabiltzen dute. Hiztegi honek aditz eta kasu batentzat probabilitate handieneko galdetzailea zein den dio.

5.1.1.3 Aurrekari azterketaren ondorio eta erabakiak

Aurrekariak aztertuta, erregeletan oinarrituriko galdera sortzaile automatiko bat eraikitzea erabaki da. Literaturan asko dira horrelako sistemak garatu

dituztenak eta emaitzak egokiak izan daitezkeela uste da.

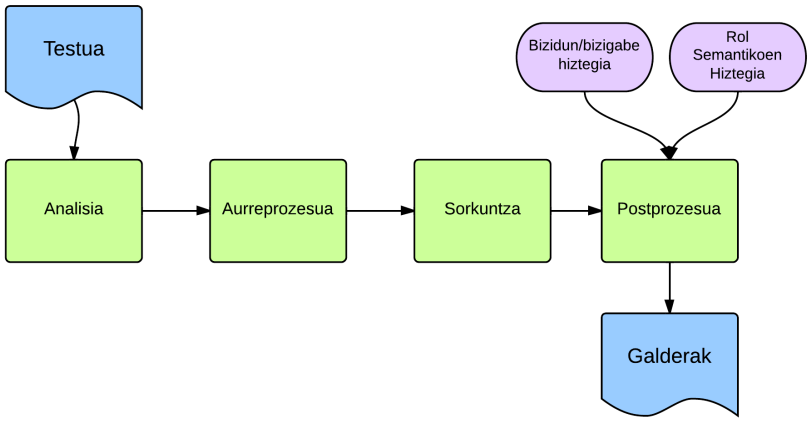
QG-Ixati sisteman erabilitako galdetzaile desanbiguazio metodoak interesgarritzat jo dira. Gainera, proiektua garatzen ari den ikerketa taldearen barruan erreminta horiek eskura daudela aprobetxatuz eta emaitza onak frogatu dituztela ikusirik garatu behar den galdera sortzailean erabiltzeko erabakia hartu da.

Bestalde galderak sortzeko orduan chunkak erabiltzetik dependentziak erabiltzeko pasa nahi da, honek emaitzak hobetuko dituelakoan. Amaitzeko *Heilman M. et al.*ek aipatzen dituzten sinplifikazio metodoak emaitza onak ematen dituztela ikusirik, hortik urrats batzuk ere jorratu nahi dira.

NuERCB sistema galdera sortzailean integratzea ere interesgarritzat jo da. Sistemarako oso interesgarria izan daitekeen arren proiekturako ez dela hain interesgarria uste dugu. Garatu nahi den proiektua ikerketara orientatu izatea nahi da eta NuERCB sistema integratzeak denbora ugari kontsumitzeaz gain ez lukeela ezer berririk aportatuko uste da. Arrazoi hauek medio, etorkizuneko lan bezala utzi da.

Aplikazioaren izenari dagokionez QG-Ixati Ixatin oinarrituriko galdera sortzailea denez, gurea QG-Malti bezala izendatu da, Maltixa analizatzaile sintaktikoan oinarrituriko galdera sortzailea delako.

5.1.2 Aplikazioaren egitura



Irudia 5.3: Galdera sortzailearen egitura

Aplikazioaren egitura orokor bat aurkezten da 5.3 irudian. Bertan lau fase ezberdintzen dira, analisia, aurreprozesua, sorkuntza eta postprozesua. Analisian testua linguistikoki analizatuko da. Ondoren, aurreprozesuan galdera eraikitzen hasi aurretik egin beharreko guztiak egingo dira. Hala nola, analisiak sorturiko fitxategiak irakurri, esaldiak sinplifikatu eta esaldi desegokiak filtratuko dira. Sorkuntza fasean, galdera eraikiko da, egin beharreko transformazioak eginez. Bertan kasuaren arabera posible diren galdetzaile zerrenda bat sortuko da. Azkenik, postprozesuan sorkuntzak utzi dituen galdetzaile posibleak murrizteko teknika ezberdinak erabiliko dira. Ahal den neurrian galdetzaile bakarra uzteko asmoz.

5.1.3 Analisi linguistikoa

Analisi fasean testua linguistikoki analizatzen da.

Lehenik testua esalditan banatzen da. Honetarako egoera finituko teknologiak erabiltzen dira. esaldiak banatzean zehaztasun handiagoa lortzeko moduak dauden arren, bertara esfortzu gehiegi ez dedikatzea erabaki da, etorkizun batean analizatzaileak berak integratuko duelako esaldiak banatzeko metodo bat.

Esaldiak analizatzeko erabili den analizatzailea Ixa ikerketa taldeko Maltixa analizatzaile sintaktikoa da. 2.1.4.2 atalean azaltzen den bezala, honek analisi sintaktikoaz gain, analisi morfologikoa, eta datu semantiko batzuk ere integratzen ditu. Analizatzaile honek analisi oso bat burutzen du, hau da, chunketatik haratagoko analisi bat ematen du, dependentzietan oinarritua. Analisi guzti hau 2.3.2 atalean azaltzen den *Computational Natural Language Learning(CoNLL)* formatuan adierazten du.

5.1.4 Aurreprozesua

Aurreprozesuan galdera sortu aurretik egin beharrekoak burutuko dira. Bertan analisiak sorturiko fitxategiak irakurtzeaz gain, esaldiak sinplifikatzen dira. Bestalde galderak sortzeko patroi desegokiak dituzten esaldiak filtratzen dira teknika ezberdinak erabiliz.

5.1.4.1 Sinplifikazioa

Esaldiak sinplifikatzeko bi metodo ezberdin erabili dira, batetik esaldi hasieran agertzen diren lokailuak kentzen dira eta bestetik juntadurazko lokailudun esaldiak bi esaldi sinpleagotan banatzen dira.

Esaldi hasierako lokailuak

Erabili den lehen sinplifikazio metodoa, esaldi hasieran ager daitezkeen lokailuak kentzea du helburu. Lokailuak aurreko esaldiarekiko erreferentzia bat izan ohi dira. Hori dela eta, galdera batean agertzean zentzua galtzen dute gehienetan. Horregatik emaitzak hobetzeko asmoz hauek hasieratik kentzea erabaki da.

Jatorrizko esaldia: Hori dela eta, Mikelek baloi bat oparitu zion Josebari.
Sinplifikazio gabeko galdera: Zer oparitu zion Mikelek Josebari, hori dela eta?

Esaldi sinplifikatua: Mikelek baloi bat oparitu zion Josebari
Galdera: Zer oparitu zion Mikelek Josebari?

Irudia 5.4: Lokailu sinplifikazioaren adibide bat

Sinplifikatu gabeko galderak zentzua galtzen du, testuingurua falta delako, 5.4 adibidean ikus daiteken moduan. Sinplifikazioan *hori dela eta* lokailua kentzen da esalditik, eta hau eginda, galdera zentzuzkoa bihurtzen da. Hau kentzeko azpikategoriatzat LOT_LOK etiketa duten nodoak aukeratzen dira.

Juntadurazko lokailuak

Juntadurazko lokailuak, bi perpaus nagusi lotzen dituzten lokailuak dira. Lotzen dituzten bi perpausak perpaus nagusiak direnez, hauek zentzu osoa dute, eta beraz banatu daitezke hauen esanahia eta zentzua mantenduz. Hona hemen adibide bat:

Jatorrizko esaldia: Mikel hondartzara joan zen eta Joseba etxean gelditu zen.
Galdera: Nor joan zen hondartzara eta Joseba etxean gelditu zen?

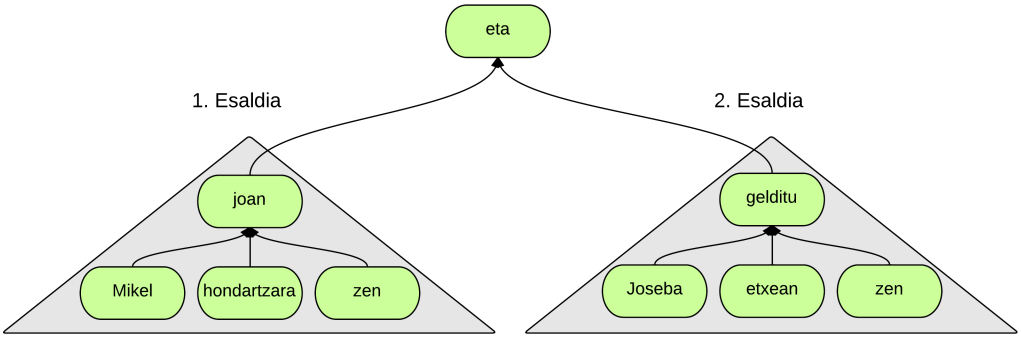
Esaldia sinplifikatua 1: Mikel hondartzara joan zen.
Esaldia sinplifikatua 2: Joseba etxean gelditu zen.
Galdera 1: Nor joan zen hondartzara?
Galdera 2: Nor gelditu zen etxean?

Irudia 5.5: Juntadurazko lokailuen sinplifikazio adibide bat

5.5 irudian ikus daitekeen moduan esaldian bi azpiesaldi nabarmentzen dira. Sinplifikazio hau egingo ez balitz esaldia gramatikalki zuzena baina zentzugabea izango litzateke. Sinplifikazioa burutu ondoren bi esaldi zuzen eta zentzudun lortzen dira eta hauen inguruan sortzen diren galderek zentzu

osoa mantentzen dute.

Sinplifikazio hau egiteko, analisi sintaktikoaren ondorioz esaldiak duen



Irudia 5.6: Juntadurazko lokailuak kentzeko zuhaitz errepresentazioa

zuhaitz errepresentazioaz baliatzen da programa. 5.6 irudian ikus daitekeen bezala, kendu nahi den juntadurazko lokailua zuhaitzaren erroan agertzen da. Lehenik juntagailuzko lokailuaren nodoaren azpian bi nodo soilik daudela konprobatzen da. Behin hau ziurtatuta erroa kendu eta bi zuhaitz independente lortzen dira, zuhaitz bakoitzak esaldi independente bat errepresentatuko duelarik.

5.1.4.2 Esaldi desegokiak filtratzeko teknikak

Kasu batzuetan esaldi bat albora utzi beharra dago, bai sortu behar diren galderetarako egokia ez delako, bai analisia ez delako ongi burutu. Kasu hauek ahalik eta modu goiztiarrean baztertzeko estrategia ezberdinak inplementatu dira. Hauek jarraian aurkezten dira:

Erro bat baino gehiagoko zuhaitzak
Batzuetan analizatzaile sintaktikoak erro bat baino gehiago duten zuhaitzak sortzen ditu. Esaldiz kanpoko erreferentzia batzuk daudelako(hauetariko batzuk konpontzen dira lokailuen sinplifikazioarekin), zein analizatzailearen inplementazioan akats batzuk daudelako sor daitezke zuhaitz hauek. Maltixa oraindik garapen prozesuan dagoela kontuan izanik ulertzekoak dira errore hauek. Esaldi hauek zuzenean baztertzen dira, aurreragoko pausoetan akatsak besterik ez baitituzte sortuko. Jarraian honen adibide bat ematen da. Bertan *dira* eta ”,”nodoak erro bezala zehazten dira analisian.

Lurreko kontinenteak etengabe ari dira mugitzen, elkartzen eta urruntzen , lurrazalaren azpian dagoen mantu likidoaren konbekzio korronteekbultzatuta.

Aditz nagusi bat esaldian
Esaldi batzuk konplexuegiak dira sorkuntza faserako. Konplexutasun hau askotan esaldian aditz nagusi ugari daudelako sortu ohi da, esaldiak konposatuak direlarik. Beste batzuetan esaldi arraroak agertu ohi dira, analisi oker edo sinplifikazio desegoki bat burutu delako, hauetan aditz nagusirik gabeko esaldiak ere agertu ohi dira. Fenomeno hauek baztertzeko asmotan esaldian egon daitezkeen aditz nagusi kopurua batera mugatu da. Aditz nagusi bakarra ez duten esaldiak baztertzen dira. Hona hemen baztertutako liratekeen adibide pare bat.

Etxera joan da aterkia ahaztu zaiolako.
Lasterketa irabazi zuen, nahiz eta bigarren lasterkariak hurbiletik jarraitu zion.

Zuhaitzaren erroa
Esaldi baten analisia egokia izateko bere erroan aditz bat edo juntagailu bat egon behar du. Hau ez da beti betetzen aurreprozesura iristen diren analisietan. Horregatik, baldintza horiek betetzen ez dituzten esaldiak baztertzen dira.

Izen bezala analizaturiko aditz nagusiak

Beste batzuetan, erroan agertzen den nodoa aditza da, baina izen portaera bat du. Horrelako aditzek aditz kategoria aurkezten dute, baina informazio morfologikoan izen atributuak dituzte. Morfologia aldetik egoki analizatuak izan diren arren, sintaxi aldetik esaldia gaizki analizatu den seinale bat da, erroan dagoen aditz nagusiak ezin baitu izen portaera bat izan. Beraz, kasu hauek detektatu eta baztertzen dira. Adibidean *galdutakoan* aditza zehaztu du analizatzaileak esaldiaren erro bezala eta ez *egiten* aditza.

Azpialdean berotutako likidoak gora egiten zuen dentsitatea galdutakoan.

Erlatibozko perpausak
Erlatibozko esaldiek arazoak sortzen dituzte galdera sorkuntzan. Kasu batzuetan ezinezkoa da hauekin galdera koherente bat sortzea, izatez galdera

bat gordetzen baitute bere barnean, zehar galderen kasuan adibidez. Horregatik baztertu egiten dira. Jarraian iragazki honek baztertuko lituzkeen pare bat esaldi erakusten dira.

Mikelek ekarritako bazkaria oso ona zen. Mikelek bazkaria ona zegoen jakin nahi du.

5.1.5 Sorkuntza

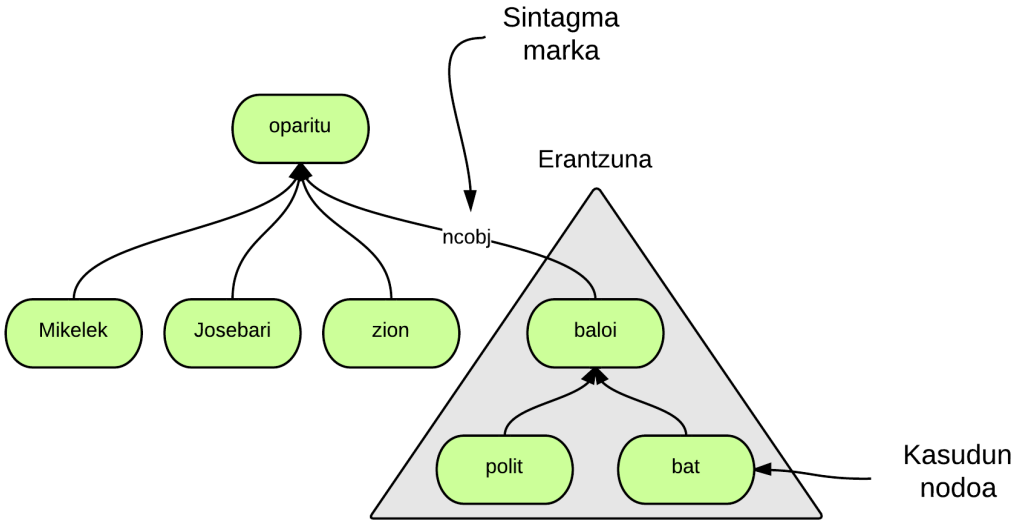
Sorkuntza fasean galdera sorkuntzaren muina garatuko da. Bertara iristen diren galdera guztiak sortzeko prest daude, dagoeneko burutu baitira sinplifikazio eta filtro guztiak. Bertan jasotako esaldian transformazioak burutuko dira eta galdetzaile posibleak zein diren zehaztuko dira.

5.1.5.1 Erantzun hautagaia aukeratu

Galdera sortzen hasteko lehen pausoa erantzuna izango den azpizuhaitza aukeratzea da. Aukeraketa hau sintagmaren kasuaren arabera burutzen da. Galderak sortzeko erabili diren kasuak honakoak dira: absolutiboa, ergatiboa, ablatiboa, adlatiboa eta inesiboa. Kasu hauek zientzia eta teknologia corpusean burutu diren neurketen arabera izen sintagmen kasu guztien agerpenen %90a osatzen dute. Hori izan da justu kasu horiek aukeratzeko arrazoi nagusia.

Kasua duen nodoaren bilaketa zuhaitzean burutzen da, goitik behera, ezkerretik eskuinera. Behin nahi den kasua duen nodoa bilatutakoan, bilaketa gelditzen da. Baina oraindik ez da lortu erantzun osoa, bere parte den izen nodo bat baizik. Horregatik, aurkitu dugun nodotik, berriro gorantz bilaketa bat burutzen da, sintagma bat markatzen duten dependentzia bat aurkitu arte. Aurkitu den nodo honek eta bere azpikoek osatzen duten zuhaitza izango da erantzuna.
Prozesu honen zergatia ulertzeko garrantzitsua da jakitea kasu marka sintagma osoarena den arren hitz mailan detektatzen den marka bat dela. Horregatik lehenik kasua duen hitza bilatu behar da eta ondoren honen sintagma zein den detektatu. Prozesu hau hobeto ulertzeko 5.7 irudia lagungarri gerta daiteke.

Galderaren erantzuna izango den azpizuhaitza zuhaitz orokorretik askatzen da, gero erantzuna sortzeko.



Irudia 5.7: Erantzun sintagmaren bilaketa

5.1.5.2 Aditza hasierara eraman

Behin zuhaitzean galderaren parte ez diren nodoak kendu eta gero bertan transformazioak egiteko momentua da. Esaldi baten adierazpen forma eta galderazkoaren arteko ezberdintasun handienetarikoa aditzaren posizioa da. Esaldi adierazle batean aditz nagusiak bukaeran agertzeko joera du, galderazko esaldietan, aldiz, galdetzailearen ondoren agertzeko joera du. Beraz esaldi bat galdera bihurtzea nahi bada transformazio hau burutu behar da.

Honetarako lehenik esaldiko aditz nagusia identifikatu behar da. 5.7 irudiari erreparatzen badiogu aditza erroan agertzen dela ikus daiteke. Bestalde, aditzaren laguntzailea *zion* bere azpiko mailan agertzen dela ikus daiteke. Beraz, aditza identifikatzeko erroa eta bere azpian egon daitezkeen aditz laguntzaileak hartzen dira.

Behin aditz nagusia osatzen duten nodoak identifikatuta, esaldiaren ordena aldatzen da, nodo hauek aurrera ekarriz.

5.1.5.3 Galdetzaile posibleak lortu

Dagoeneko esaldia bere galderazko forman aurkitzen da. Hurrengo pausoa galderaren hasieran agertuko den galdetzaile posibleak identifikatzea da. Galderak kasuan oinarriturik sortzen direnez galdetzailea ere honen araberakoa da, hona hemen kasu bakoitzarentzat galdetzaile posibleak.

Kasua	Galdetzaile posibleak
Absolutiboa	Zer, Nor
Ergatiboa	Zerk, Nork
Adlatiboa	Nora, Norengana
Ablatiboa	Nondik, Norengandik
Inesiboa	Non, Norengan, Noiz

Taula 5.4: Kasu eta galdetzaileen arteko erlazioak

Ikus daitekeen moduan, kasu bakoitzarentzat galdetzaile bat baino gehiago ondoriozta daiteke. Honek hurrengo fasean beste estrategia batzuk inplementatu beharra ekartzen du, agertzen diren bi edo hiru galdetzaileak batera murrizten saiatzeko.

5.1.6 Postprozesua

Postprozesua dagoeneko galderak eraikita daudela burutuko da. Bere helburua, sortutako galderak hobetzea da. Galdera guztietan ematen den arazo bat galdetzaileena da. Galdera bakoitzarentzat bi edo hiru galdetzaile posible sortu dira sorkuntza fasean. Beraz, postprozesuan galdetzaile hauek desanbiguatzeko estrategia ezberdinak inplementatu dira. Estrategia hauek jarraian agertzen diren ordena berean exekutatzen dira.

5.1.6.1 Entitate bidezko desanbiguazioa

Entitate bidezko desanbiguazioa Eiherak[2] detektatutako entitate izendatuen informazioan oinarritzen da. Eiherak termino bat leku izen berezi bat edo pertsona izen berezi bat den detektatzen du. Informazio hau garrantzizkoa da galdetzailea desanbiguatzeko. Adibidez, absolutibo kasuan erantzuna pertsona izen bat dela detektatuz gero badakigu galdetzailea *nor* izango dela eta ez *zer*. Leku izen bereziekin ere, adibidez inesibo kasuan galdetzailea *non* dela determina daiteke.

Jarraian 5.4 taularen bi bertsio ezberdin erakusten dira. Bata erantzuna pertsona izen berezia dela detektatu denean geldituko litezkeen galdetzaileak eta berdina leku izen bereziekin.

[2]Informazio gehiagorako ikusi 2.1.3 atala.

Kasua	Galdetzaile posibleak
Absolutiboa	Nor
Ergatiboa	Nork
Adlatiboa	Norengana
Ablatiboa	Norengandik
Inesiboa	Norengan

Taula 5.5: Kasu eta galdetzaileen arteko erlazioak, pertsona izen berezia detektatuta

Kasua	Galdetzaile posibleak
Absolutiboa	Zer
Ergatiboa	Zerk
Adlatiboa	Nora
Ablatiboa	Nondik
Inesiboa	Non

Taula 5.6: Kasu eta galdetzaileen arteko erlazioak, leku izen berezia detektatuta

Ikus daitekeenez entitate izendatu bat detektatzen den bakoitzean galdetzaile bat aukera dezakegu kasu guztietan, beraz estrategia egokia da galdetzaile desanbiguaziorako. Zoritxarrez, testuetako entitate izendatu kopurua txikia da, beraz estrategia honek funtzionatzen duen arren, beste estrategia batzuk ere inplementatu behar dira honekin batera.

5.1.6.2 Bizidun/bizigabe hiztegiaren bidezko desanbiguazioa

Galdetzaile desanbiguaziorako inplementatu den bigarren estrategia bizidun/-bizigabe hiztegiaren[3] erabilera da. Hiztegi honen bitartez izen bat bizidun edo bizigabea den jakin daiteke. Sintagman bilaketa bat burutu behar da bere gunea(bere hitzik garrantzizkoena) lortzeko, hitz hau izango baita hiztegian bilatuko dena. Sintagma bateko izen nagusia bere zuhaitzean gorena agertzen den izena da. Bilaturiko izen hori erabiltzen da hiztegiarekin konparatzeko. Jarraian bi taula erakusten dira izena bizidun edo bizigabea dela detektatu ondoren dauden galdetzaile posibleekin.

[3]Informazio gehiagorako ikusi 2.2.2 atala.

Kasua	Galdetzaile posibleak
Absolutiboa	Nor
Ergatiboa	Nork
Adlatiboa	, Norengana
Ablatiboa	Norengandik
Inesiboa	Norengan

Taula 5.7: Kasu eta galdetzaileen arteko erlazioak, izena biziduna denean

Kasua	Galdetzaile posibleak
Absolutiboa	Zer
Ergatiboa	Zerk
Adlatiboa	Nora
Ablatiboa	Nondik
Inesiboa	Non, Noiz

Taula 5.8: Kasu eta galdetzaileen arteko erlazioak, izena bizigabea denean

Kasu gehienetan galdetzaileak batera murrizten dira, inesibo bizigabean izan ezik bertan bi aukera baitaude oraindik. Bizidun/bizigabe hiztegiak emaitza egokiak ematen ditu kasu gehienetan. Zoritxarrez hiztegian ez daude izen posible guztiak eta batzuk anbiguo bezala etiketatuta agertzen dira, kasuan-kasuan bizidun edo bizigabe izan daitezkeelako edo ez dagoelako bizitasuna zehazterik. Kasu horietan hirugarren estrategia aplikatu beharko da.

5.1.6.3 Rol semantikoen bidezko desanbiguazioa

Azken estrategia hau, bizidun/bizigabe hiztegiarekin izenaren bizitasuna determinatu ezin izan denean soilik exekutatzen da. Rol semantikoen hiztegian[4] asko erabilitako zenbait aditzen inguruan agertzen diren kasuen rol semantiko ohikoena adierazten da. Rol semantikorekin batera galdetzaile posibleena ere adierazten da. Hiztegian bilaketa bat burutuz determinatzen da galdetzailea.

Hiru estrategiak exekutatu ondoren, kasu batzuetan galdetzaile posible bat baino gehiago gelditzen da. Kasu hauetan bi edo hiru galdetzaileak bere horretan uzten dira galderan, etorkizunean hiztegi hau zabalagoa bihurtuko den itxaropenean.

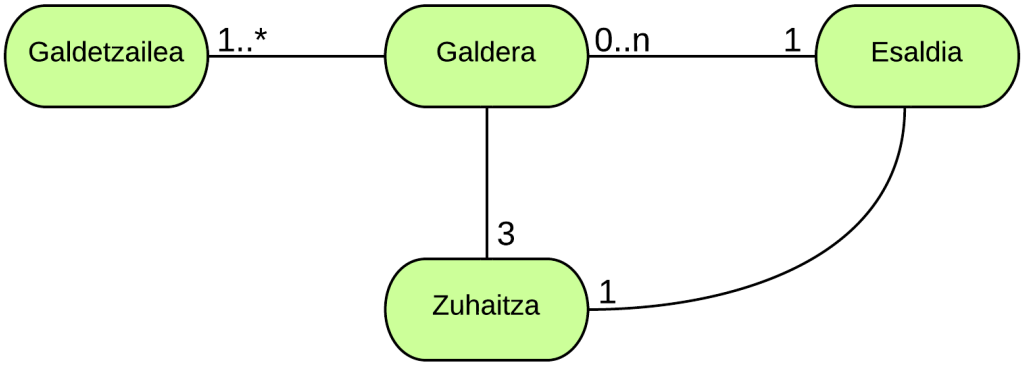
[4]Informazio gehiago lortzeko ikusi 2.2.1 atala.

5.2 Diseinua

Jarraian galdera sortzailearen inguruko zeresanak azalduko dira programazio ikuspegi batetik. Hau inplementatzeko erabili diren estrategiak, egiturak, eta beren metodo esanguratsuenak azalduko dira.

5.2.1 Domeinu eredua

Atal honetan domeinua errepresentatzeko erabili diren egiturez zein beraien arteko erlazioez hitz egingo da. 5.8 irudian galdera sortzailearen domeinu



Irudia 5.8: Galdera sortzailearen domeinu eredua

eredua erakusten da. Bertan lau objektu ezberdin ikus daitezke. Hauek jarraian azaltzen dira:

5.2.1.1 Esaldia

Esaldia objektuak galdera sortzaileak jasotzen dituen esaldi bakoitza errepresentatzen du. Bere barnean esaldiaren *Zuhaitz* egitura eta esalditik sortu diren galdera guztiak gordetzen ditu. Bertako galderak sortzeko metodo batez gain esaldia bera zein galderak modu ezberdinetan inprimatzeko aukerak ematen ditu.

5.2.1.2 Galdera

Galdera esaldietan sortu den galdera bakoitza errepresentatzeko egitura da. Bere barnean honako datuak gordetzen ditu:

- **3 zuhaitz**
  Hiru zuhaitz egitura ezberdin gordetzen dira bere barnean. Batetik, ja-

torriz pasa zaion zuhaitz osoa. Bestetik galdera osatuko duten nodoen zuhaitza. Eta azkenik erantzuna osatuko duten nodoen zuhaitza.

- **kasua**
  Galdera sortzeko erabiliko den kasua da eta erantzuna bilatu zein galdetzaileak aukeratzeko erabiltzen da.

- **galdetzaileak**
  Galderaren galdetzaile posible guztiak gordetzen dira zerrenda honetan.

Galdera klaseak dituen metodoen artean galderak sortzen laguntzeko metodoak zein hauek inprimatzekoak aurkitzen dira. Bestalde, galdetzaileak sortzeko eta murrizteko metodo ezberdinak ere inplementatuak ditu.

5.2.1.3 Galdetzailea

Galdetzailea objektuak galderaren barruan dauden galdetzaile posible bakoitza errepresentatzen du. Gordetzen duen informazioa galdetzailearen izena (non, nor, zer...) eta bere bizitasunaren inguruko informazioak osatzen dute. Inplementatuta dituen metodoak berau inprimatzeko zein bizitasunaren inguruko informazioa jakiteko metodoak dira.

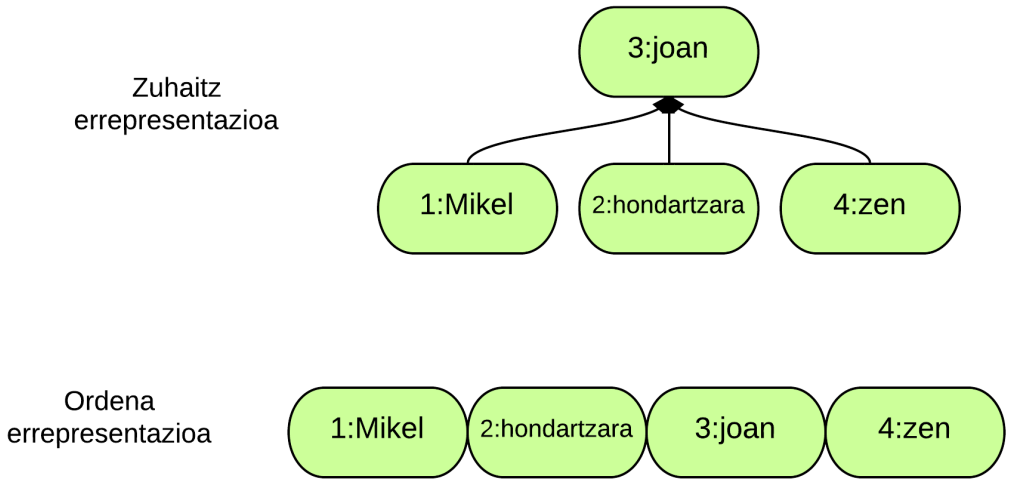
5.2.1.4 Zuhaitza

Zuhaitz egitura analizatzaileak emandako datu guztiak gordetzen dituen egitura da. Nodo bakoitza *zuhaitz* objektu bat da, eta bere guraso eta umeen erreferentziak gordetzen ditu. Erreferentzia sistema horren bitartez zuhaitz egitura bat osatzen da. Nodo bat erroa izango da, gurasoaren erreferentziarik ez duenean, eta hostoa izango da umerik ez duenean.

Nodoek beren zuhaitz egitura errepresentatzeko erreferentziez gain informazio gehiago dute. Hala nola, errepresentatzen duten hitza, lema, informazio morfologikoa, kategoria, azpikategoria eta dependentzia.

Bestalde, esaldiko terminoen ordena informazioa gordetzeko atributu bat du nodo bakoitzak. Atributu hau beharrezkoa da transformazio batzuk egiteko, zuhaitz errepresentazioak ez baitu ordenarik adierazten. Adibidez esaldiaren aditza hasierara eramateko transformazioa burutzeko posizio atributuak aldatzen dira. Honekin esaldiek bi errepresentazio dituztela esan daiteke. Batetik zuhaitz errepresentazioa, analisiaren informazioa jakiten lagun dezakeena. Bestetik, ordea, zerrenda ordenatu moduko errepresentazio bat ere sortu da, nodoak esaldian zetozen ordenan gordetzeko asmoarekin. Bi

errepresentazioen artean aldaketak errazteko asmoz, zuhaitz bat ordena errepresentazioan lortzeko funtzio bat inplementatu da. 5.9 irudian esaldi baten bi errepresentazioak ikus daitezke adibide gisa.



Irudia 5.9: Esaldi bat bere bi errepresentazioetan

Zuhaitzekin lana egitea errazteko zenbait funtzio lagungarri inplementatu dira, edozein zuhaitz generikotan inplementatu ohi direnak:

- **hostoaDa()**
  Nodoa hostoa den dio, hau da ume kopurua 0 ote den.
- **erroaDa()**
  Nodoa zuhaitzaren erroa den dio, hau da gurasorik ez duen.
- **banatuGurasotik()**
  Nodo bat bere gurasotik banatzeko balio du. Gurasotik umerako erreferentzia zein umetik gurasorakoa ezabatzen ditu.
- **lortuZerrendaModuan()**
  Batzuetan zuhaitz bateko nodo guztiak zeharkatu behar direnean, guztiei eraldaketa bat egiteko adibidez, zuhaitz egitura desegokia gerta daiteke. Kasu horietan errazagoa da zuhaitzaren nodo guztien zerrenda batekin lan egitea. Funtzio honek horretarako balio du, hau da zuhaitz egitura bat nodo zerrenda batean bihurtzen du.
- **lortuMailaka()**
  Honako funtzio hau *lortuZerrendaModuan()* funtzioaren eraldatze bat

da. Bere berdina egiten du, baina ordenazio berezi batekin. Nodoak zuhaitzean dauden mailetan itzultzen ditu, hau da goitik behera, ezkerretik eskuinera. Ordenazio hau interesgarria da zabalerako bilaketak burutzeko, kasu batzuetan baldintza batzuk betetzen dituen nodo gorena behar baita.

Kontsulta sistema

Zuhaitzarekiko eragiketak errazteko kontsulta sistema bat inplementatu da. Honek baldintza berezi batzuk dituen nodoak bilatzeko aukera ematen du behin eta berriz bilaketa ezberdinak inplementatzen ibili beharrean. Sistemaren oinarria *nodoakBaldintzakBetetzenDitu()* izeneko funtzio batean dago. Nodoak argumentu bezala zehaztutako baldintzak betetzen dituen edo ez itzultzen du. Honako argumentuak jasotzen ditu:

- **lema**
  Nodoaren lemak bete behar duen adierazpen erregularra adierazten du. Honek lema konkretu batzuk lortzeko aukera ematen du. Hala nola, "a"hizkiaz hasten direnak, edo luzera zehatz bat dutenak, adibide batzuk jartzearren.
- **kategoria**
  Nodoaren kategoriak bete beharreko adierazpen erregularra.
- **azpikategoria**
  Nodoaren azpikategoriak bete beharreko adierazpen erregularra.
- **informazio morfologikoa**
  Informazio morfologikoa zehazteko zerrenda bat.
- **informazio morfologiko kopurua**
  Aurreko zerrendatik zenbat elementuk bete behar duten.
- **dependentzia**
  Nodoaren dependentziak bete beharreko adierazpen erregularra.

Hona hemen erabilpenaren zenbait adibide:

```
1 nodoakBaldintzakBetetzenDitu(".*", ".*", ".*",
     {"ERL:KONPL","ERL:ZHG","ERL:ERLT,}, 0, ".*")
2 nodoakBaldintzakBetetzenDitu(".*", ".*", ".*",
     {"KAS:ABS","NUM:S"}, 2, ".*")
3 nodoakBaldintzakBetetzenDitu(".*", "ADT|ADI", ".*",
     {"KAS:ABS","NUM:S"}, 0, ".*")
```

Lehen adibidean nodoak ez duela ez konpletiboa(KONPL) ez zehar galdera(ZHG) eta ez erlatibozkoa(ERLT) izan behar esaten da. Horretarako konparatu nahi den informazio morfologikoa pasatzen da argumentu gisa eta horietatik 0k bete behar dutela esaten da, hau da ez duela inork bete behar. Bigarrenen adibidean nodoak absolutibo(ABS) eta singularra(S) izan behar duela esaten da, biak batera. Azkenik hirugarren adibidean nodoak edo aditz trinkoa(ADT) edo aditz nagusia(ADI) izateaz gain ez duela absolutiboa(ABS) eta singularra(S) izan behar esaten da.

nodoakBaldintzakBetetzenDitu() funtzioaren gainean beste zenbait funtzio inplementatu dira. Hauek jarraian aurkezten dira:

- **lortuBaldintzadunNodoGorena()**
  Zuhaitzean baldintzak betetzen dituen nodo gorena lortzen du, goitik behera, ezkerretik eskuinera.
- **lortuBaldintzadunNodoGuztiak()**
  Baldintzak betetzen dituzten nodo guztiak lortzen ditu.
- **BaldintzadunNodoKopurua()**
  Baldintzak betetzen dituzten nodo kopurua itzultzen du.
- **existitzenDaBaldintzadunNodoa()**
  Baldintzak betetzen dituen nodorik existitzen den dio
- **existitzenDaBaldintzadunNodoaTartean()**
  Zuhaitzean adar berean dauden bi nodoren artean baldintzadun nodorik existitzen ote den dio.

Funtzio lagungarri guzti hauek malgutasun handia ematen dute zuhaitzean burutuko diren eragiketak ahalik eta modu sinple eta errazean burutzeko. Kodean zehar funtzio guzti hauek behin eta berriz erabili dira bilaketa zein egiaztapen ezberdinak egiteko.

5.2.2 Bizidun/bizigabe hiztegirako klaseak

Bizidun/bizigabe hiztegiarekiko abstrakzio geruza bat inplementatu da. Honen helburua hiztegiaren erabilera errazteaz gain bertan gerta daitezkeen aldaketekiko mendekotasuna gutxitzea da. Horrela hiztegiaren egitura aldatuko balitz, geruza hau berrinplementatuko beharko litzateke bakarrik eta ez programa osoa.

Geruza honetan honako funtzio lagungarriak inplementatu dira:

- **kargatu()**
  Hiztegia memorian kargatzeko balio du.
- **bizidunaDa()**
  Izen bat biziduna den dio.
- **bizigabeaDa()**
  Izen bat bizigabea den dio.
- **hiztegianDago()**
  Izen bat hiztegian dagoen dio.

5.2.3 Rol semantikoen hiztegirako klaseak

Rol semantikoen hiztegiarekiko ere abstrakzio geruza bat inplementatu da. Honen helburua hiztegiaren erabilera errazteaz gain bertan gerta daitezkeen aldaketekiko mendekotasuna gutxitzea da. Horrela hiztegiaren egitura aldatuko balitz, geruza hau berrinplementatuko beharko litzateke bakarrik eta ez programa osoa.

Geruza honetan honako funtzio lagungarriak inplementatu dira:

- **kargatu()**
  Hiztegia memorian kargatzeko balio du.
- **lortu galdetzailea()**
  Aditz eta kasu bat emanda bere galdetzaile ohikoena itzultzen du.
- **existitzenDa()**
  Bilatu nahi den aditza emanda, hau hiztegian jasota dagoen edo ez dio. Bestalde, kasu bat ere zehaztu daiteke, aditz konkretu horren sarreran kasu hau jasota dagoen edo ez esan dezan.

5.3 Ebaluazioa

Jarraian QG-Maltiren funtzionamendua ebaluatzeko garatu diren bi ebaluazioen inguruko azalpenak ematen dira. Bi ebaluazio garatu dira, bata proiektua garapen fasean zegoela, guztiz bukatu eta zuzendu gabe. Ebaluazio honen helburua bikoitza zen. Batetik QG-Malti eta QG-Ixatiren areko konparaketa bat burutzea. *SEPLN*rako argitaratu den artikuluan[9] ikus daiteke zehatzago. Bestetik proiektuaren garapen faseko ebaluazio bat lortzea. Bigarrena garapen fasea bukatu eta gero burutu da. Ez da ebaluazio oso bat izan genituen denbora murriztapenengatik, baina sistemak izan dituen hobekuntzak aurreikusteko balio izan du.

5.3.1 Garapen faseko ebaluazioa

Ebaluazio honetan QG-Malti eta QG-Ixatiren arteko konparaketa bat aurkezten da. Hau burutzeko derrigorrezko bigarren hezkuntzarako (DBH) testuak zein hizkuntzak irakasten espezializaturiko corpus bateko testuak erabili dira. Ebaluazioan hizkuntzalari batek galderen zuzentasun sintaktikoa neurtzeaz gain galdetzaileen egokitasuna ere neurtu du.

5.3.1.1 Testuak

Bi testu multzo erabili dira ebaluazioa burutzeko. Lehenengo multzoa zientzia eta teknologia arloko bost testuz osatua dago. Irakaskuntza materialak sortzen dituen aditu batek testu hauek DBH mailako testu bezala sailkatu ditu. Testu hauen gaiak, kontinentea, planeta, saguzarrak, artikoa eta ordenagailuak dira. Testu guztiek luzera antzekoa dute (35 esaldi inguru), guztira 176 esaldi dituztelarik. Bertako esaldien %60ak absolutibo kasu markak dituzte, %11ak ergatibo markak, %16ak inesibo markak eta beste kasu guztiek %4a baino gutxiago osatzen dute. Galdera sortzaileek sortzen dituzten markak esaldien %90ean aurki daitezke.

Bigarren testu multzoa euskara ikasten laguntzeko corpus espezializatu batetik lorturiko testuek osatzen dute. Marko europarreko tarteko mailako testuak hartu dira bertatik. Corpusean 80.000 esaldi aurki daitezke, bataz beste 13 hitzekoak.

5.3.1.2 Esperimentuak

Atal honetan bi esperimentu aurkezten dira, esperimentu bana testu multzo bakoitzeko. Esperimentu bakoitzean bi datu neurtzen dira:

Batetik, zenbat kasutan aukeratu den galdetzaile egokia. Honetarako hizkuntzalari batek etiketatu ditu bai/ez moduan galdera guztiak. Hau egiteko jatorrizko esaldia, galderaren erantzuna eta galdetzailea soilik kontuan hartu ditu.

Bestetik, galderen gramatikaltasuna ere neurtu da. Honetarako ebaluatzaileak galdera soilik kontuan hartu du. Hau ebaluatzeko ondorengo puntuazioak erabili dira:

- 4 – galdera gramatikalki zuzena da.
- 3 – galderak aldaketa txikiak behar ditu zuzen egoteko.

- 2 – galderak aldaketa handiak behar ditu zuzen egoteko.
- 1 – galdera ez da gramatikala eta ez dago zuzentzerik.

5.3.1.3 Zientzia eta teknologiako testuen esperimentua

Esperimentu honetan bi galdera sortzaileek sor ditzaketen galdera posible guztiak sortu dituzte. QG-Maltik 112 galdera sortu ditu eta QG-Ixatik 81.

	Galdetzailea	**Gramatikaltasuna**
QG-Malti	75%	57'1%
QG-Ixati	87'6%	59'3%

Taula 5.9: Gramatikaltasun eta galdetzaile ebaluazioa

5.9 taulan ebaluazioaren emaitzak ikus daitezke galdetzaile zein gramatikaltasunari dagokionez. Gramatikaltasunaren zutabean galdera gramatikalak zein aldaketa gutxi behar zituztenak hartu dira kontuan.

Bertan ikus daitekeenez, QG-Ixatik emaitza hobeak ematen ditu doitasunari dagokionez. Hala ere, QG-Maltik galdera on gehiago sortzen ditu besteak baino kopuruz. QG-Maltik guztira 64 galdera zuzen sortzen ditu QG-Ixatiren 48ren alboan.

QG-Malti eta QG-Ixatiren antzekotasunen ondorioz galderak sortzen hautagai berdinetik hasten dira askotan. 112 eta 81 galderetatik, 50 galderak hautagai komuna izan dute. Kopuru hau txikia da sistemen arteko antzekotasuna kontuan izanik. Honen arrazoia erantzun hautagaia aukeratzeko metodoan dago. QG-Malti sistemak kasua duen nodo gorena hartzen duen moduan, QG-Ixatik ez du hau horrela egiten. Honek kasua duen nodo bat entitate izendatua bada lehentasuna ematen dio.

Bai gramatikaltasun zein galdetzaileen egokitasuna neurtu da hautagai komuneko galderetan zein hautagai ezberdinekoetan. Datu hauek 5.10 taulan ikus daitezke.

Komunean dauden galderez aparte, QG-Maltik 62 esaldi gehiago aukeratzen ditu galderak sortzeko, QG-Ixatik aldiz, 31. 5.10 taulari erreparatuz gero, adierazgarria da QG-Maltiren hobekuntza komunean ez diren hautagaietan. Bestalde QG-Ixatik emaitza okerragoak lortzen ditu komunean ez

	Hautagai komuna		Hautagai ez komuna	
	Galdetzailea	**Gramatikaltasuna**	**Galdetzailea**	**Gramatikaltasuna**
QG-Malti	76%	54%	72,6%	59,7%
QG-Ixati	88%	66%	87,1%	48,4%

Taula 5.10: Ebaluazioaren emaitzak, hautagai komuna eta ez komunarekin

	QG-Malti		QG-Ixati	
	Galdetzailea	**Gramatikaltasuna**	**Galdetzailea**	**Gramatikaltasuna**
ABS	70%	60%	85%	85%
ERG	**95%**	45%	**80%**	65%
INE	60%	85%	70%	85%
ADL	70%	**45%**	85%	**35%**
ABL	50%	**50%**	55%	**40%**

Taula 5.11: Ehunekoak kasuko (20 galdera kasu bakoitzeko)

dituzten nodoei erreparatuz, komunean daudenekin konparatuz.

5.3.1.4 Irakaskuntzako testuen esperimentua

Irakaskuntzako testuekin burutu den esperimentuaren helburua kasu marken kopuruek emaitzetan izan dezaketen eragina neurtzea da. Horretarako hartu den lagina kasu bakoitzeko 20 galderakoa izan da. Beraz, 5 kasu ezberdin ebaluatu direnez, guztira 100 galderako lagina hartu da sistema bakoitzetik.

5.11 taulan ebaluazioaren emaitzak ikus daitezke, kasuen arabera banatuta. Orokorrean, bi sistemek emaitza hobeak lortzen dituzte absolutibo zein inesibo kasuetan. QG-Ixatik orokorrean emaitza hobeak lortzen ditu QG-Maltirekin konparatuz. Salbuespen bakarra ergatibo kasua da. Adierazgarria da bi sistemen arteko absolutibo kasuen ezberdintasuna, bertan QG-Ixatik %85a lortzen du QG-Maltiren %60aren alboan. Honen arrazoi nagusia analizatzailean dagoela uste da. Maltixa analizatzaileak absolutibo kasua duten sintagmekin okerrago dabilen susmoa dugu. Ez da froga sakonik burutu, baina eskuzko berrikuspenek hori erakusten dute.

Datu orokorrak eta hautagai komuneko galderak bakarrik kontuan harturik lorturiko emaitzak 5.12 taulan ikus daitezke. Orokorrean QG-Ixatik emaitza hobeak lortzen ditu. Hala ere adierazgarria da hautagai komunak soilik hartuz lortzen diren emaitzak. Kasu honetan QG-Ixati okertu egiten

	Orokorra		Hautagai komunak	
	Wh-word	**Grammar**	**Wh-word**	**Grammar**
QG-Malti	69%	57%	70,2%	63,8%
QG-Ixati	75%	62%	68,1%	59,6%

Taula 5.12: Hautagai komun eta ez komunak

da (%62tik %59,6ra), baina QG-Malti hobetu egiten da (%57tik %63,8ra).

5.3.1.5 Errore analisia

Datuak ikusirik, QG-Malti sistema hobetzen jarraitzeko estrategia ezberdinak identifikatu dira. Batetik sintagmak markatzen dituzten dependentzia berriak identifikatu dira. Dependentzia hauek identifikatu gabe zeuden. Horregatik esaldi batzuetan erantzun bezala esaldi osoa hartzen zen, sintagma markaren bila hasi eta erroraino iristen baitzen bilaketa.

Beste alde batetik erlatibozko esaldi guztiak gaizki sortzen zirela ikusi da. Hori dela medio, esaldi hauek hasieratik baztertzea erabaki da, QG-Ixatik egiten duen modu berean.

Azkenik, jatorrizko esaldi batzuk ez gramatikalak ziren izatez. Jatorrizko esaldia ez gramatikal izanda ia ezinezkoa da galdera gramatikal bat lortzea. Horregatik ez gramatikaltasun hau detektatzeko bi estrategia inplementatu dira, esaldi hauek zuzenean baztertu ahal izateko. Estrategia hauek analisi sintaktikoan patroi arraroak detektatzean oinarritzen dira. Esaldi bat ez gramatikala denean analisi zuhaitzaren erroan aditzik ez egoteko joera bat detektatu da. Bestalde erroko aditzak, esaldiko aditz nagusiak ez diren kasu batzuk identifikatu dira. Hauetan erroan dagoen aditza izen bihurtua da, eta beraz, izen baten informazio morfologikoa du. Erroan izen bihurturiko aditz bat egotea analisi sintaktikoa gaizki burutu den seinale bat da. Bi joera hauek detektatzeko eta baztertzeko filtroak inplementatu dira. Honen adibidea 5.1.4.2 atalean ikus daiteke.

5.3.1.6 Ondorioak

Ebaluazioa ikusirik QG-Ixatik emaitza hobeak ematen dituela ikusi da. Hala ere emaitza hauek ez dira nabarmen hobeak. QG-Malti oraindik garapenean

dagoen sistema bat dela kontuan hartuz QG-Maltik QG-Ixatik baino emaitzak hobeak lortuko dituela aurreikusten da, behin garapen fasea bukatuta. Gainera, errore analisiaren ondorioz sistema hobetzeko estrategia ezberdinak detektatu dira. Estrategiok sisteman zuzeneko hobekuntza bat ekarriko dutela aurreikusten da.

5.3.2 Lehen pausoak bukaerako ebaluazio bati begira

Behin garapena bukatuta konpondu diren erroreek emaitzetan zein ondorio izan dituzten ikusteko ebaluazio bat prestatu da. Bertan aurretik erabili ziren zientzia eta teknologiako testu berdinak hartu dira eta bertan sistemak sortzen dituen galderen gramatikaltasuna ebaluatu da.

	Gramat. garapena bukatuta	**Gramat. garapenean**
QG-Malti	71'8%	57'1%
QG-Ixati	59'3%	59'3%

Taula 5.13: Gramatikaltasun eta galdetzaile ebaluazioa

5.13 taulan QG-Malti sistemaren prototipoaren garapena bukatu ostean dituen emaitzak ikus daitezke, garapen fasean zituenen alboan. Ikus daitekeen moduan emaitzak %15 inguru hobetu dira. Gainera, QG-Malti sistemaren emaitzak QG-Ixati sistemarenak baino hobeak direla %10 batean ikus daiteke. Bestalde sorturiko esaldi kopuruari dagokionez, QG-Ixatik 81 galdera sortu ditu eta QG-Maltik 96. Bi sistemen prototipoak bukatuta egonik QG-Malti sistemaren bai kopuru zein doitasun aldetik QG-Ixatirenak baino hobeak direla esan daiteke.

Dena den, hau ziurtatzeko, batetik, testu berri batzuen gainean burutu beharko litzateke konparaketa eta bestetik, bi ebaluatzailek burutu beharko lukete, termino erauzlearekin egin den moduan. Ebaluazioa modu honetan burutzea erabaki da denbora murriztapenengatik, baina etorkizun hurbileko lana ebaluazio sakonago bat burutzea izango da.

6 Kapitulua

Ondorioak eta etorkizuneko lana

Gaien Aurkibidea

6.1 Ondorioak

Behin proiektua bukatutzat emanda, honen inguruan gertaturiko gora-behera guztiak analizatzeko momentua da. Jarraian proiektuak izan dituen ondorioak zein ondorio pertsonalak deskribatzen dira.

6.1.1 Proiektuaren ondorioak

Proiektuan euskararako galdera sortzaile automatikoen inguruan bi ekarpen gauzatzea lortu da:

Batetik eduki aukeraketa burutzeko termino erauzleen balioa aztertu da. Azterketa honek termino erauzle sistema baten prototipo bat izan du emaitza gisa. Prototipo hau gaur egunean existitzen diren sistemekiko parean dagoela frogatu da, zenbait kasutan hauek gaindituz. Hori dela medio, etorkizunean termino erauzle baten beharra izango duten sistemetan baliagarria izango dela espero da.

Bestetik galdera eraikuntza zein galdera motaren aukeraketan azterketa bat garatu da. Honek ere prototipo baten sorrera ekarri du. Sistema honekin egin den bukaerako ebaluazioak orain arte zeuden euskararako sistemekiko doitasun aldetik hobekuntza nabari bat lortu dela ondorioztatzen du. Honekin galderen eraikuntzan dependentzietan oinarrituriko analizatzaile sintaktiko bat erabiltzeak hobekuntza nabariak sor ditzakeela aurreikusten da. Galdera kopuruari dagokionez ere emaitzak hobeak direla ikusi da. Hau sinplifikazioan burutu diren saiakeren ondorioa izan dela uste da. Hala ere, emaitza hauek sakonago aztertu beharko dira etorkizun hurbilean.

Corpusak eskuratzeko burutu den lana etorkizunean baliagarria izango dela uste da. Corpus berri bat sortu da historia arloan eta Wikipediatik corpusak biltzeko tresna baliagarri bat ere sortu da. Wikipediatik sorturiko corpusak zaratatsuak gerta daitezke oraingoz. Baina erremintan hobekuntza batzuk burutuz, aplikazio ugaritarako egokiak izan daitezkeen corpusak biltzeko gai izan daitekeela uste da.

Proiektuarekin oso erlazionaturik dagoen *hizkuntzaren prozesamendua* irakasgaian ikasleak burutu behar zuen lan batean proiekturako baliagarria izan zitekeen lan bat burutzeko ideia sortu zen. Bien arteko elkarrizketen ondorioz galdera sortzaile sistema gailu mugikorretarako integratzea erabaki zen. Lan honen ondorioz android sistema eragilerako prototipo bat sortu zen. Aplikazioak oraindik eduki nahiko genituzkeen funtzionalitate guztiak ez dituen

arren, dagoeneko testu bat idatzita bertatik galdera erantzun pareak erakusteko gai da. Interfaze aldetik hobekuntza batzuk eginez irakaskuntzarako aplikazio interesgarri bat izatera iritsi daitekeela uste da. Aplikazio honen argazki batzuk ikusi nahi izanez gero V eranskina ikus daiteke.

Azkenik, proiektuan garatu den QG-Malti sistemaren ondorioz, SEPLN-rako artikulu bat[9] argitaratzea lortu da. Bertan QG-Malti eta QG-Ixati deskribatzeaz gain bien artean gauzatu den konparaketa deskribatzen da.

Beraz bada, proiektuan planteatzen ziren helburuak guztiz burutu dira eta espero ez ziren beste bi ataza ere burutzea lortu da, gailu mugikorretarako integrazio zein artikuluaren argitalpenarekin.

6.1.2 Ondorio pertsonalak

Ondorio pertsonalei dagokionez proiektu hau garatu izana oso aberasgarria izan dela uste dut. Nire formakuntzari dagokionez, jakintza berri ugari lortu ditudala uste dut. Batez ere, hizkuntzaren prozesamendurako tresnen inguruan gauza berri asko ikasi ditut. Honen ikuspegi orokor bat lortu dudala uste dut, batetik bere alderdirik linguistikoena aztertuz galdera sortzailearekin eta, bestetik, bere alderdi estatistikoagoa jorratuz termino erauzlea garatzeko.

Gainera, proiektua ikerketa talde baten barruan garatu izanak taldeko funtzionamendua bertatik bertara ikusteko aukera eman dit. Honek ikerketarako interesa piztu dit eta nire etorkizuna ikerkuntzaren inguruan kokatzeko nahia sortu dit.

Espezialitateko ikasgaietan ikasitako gauzei etekina atera diedala uste dut. Batetik, *datu meatzaritza* zein *estatistika* irakasgaietan ikasitako metodo askok aplikazio zuzena izan dute termino erauzle eta ebaluazioan. Bestetik, *algoritmoen diseinua* irakasgaiko teknikak ere aplikatu ditut zuhaitzen zein grafoen tratamenduan. Proiektuan burututako lanak ikasitako zenbait ezagutza hobeto barneratzen lagundu didala uste dut.

Idatziriko argitalpena SEPLN-n onartua izanak, proiektua, zein bertan erabili diren teknikak gure taldetik kanpo ere baloratu direla ikustera eraman nau. Honek guztion partetik burutu den esfortzua merezi izan duela ondorioztatzera eraman nau.

Amaitzeko, proiektuaren garapenean, zailtasunak alde batera, oso eroso eta gustura sentitu naizela esan behar dut. Proiekturako zuzendariarekiko

zein lan taldearekiko harremana oso ona izan da. Zailtasun momentutan taldea gertu sentitu dut, beti laguntzeko prest eta horrek proiektuaren emaitza onak bideratu dituela uste dut.

6.2 Etorkizuneko lana

Ikerketa proiektu batek ez du bukaera finkorik. Ikerketa burutzen ari den bitartean gauza interesgarri ugari agertzen dira hobetzeko. Nahiz eta hobekuntza ugari aurreikusi proiektua bukaera bat jarri behar izan zaio. Hau horrela, etorkizunean burutzeko interesgarriak izan daitezkeen ataza batzuk identifikatu dira. Hauek jarraian zerrendatzen dira.

- Garatu diren bi moduluak integratzeko metodoak aztertzea eta integratzea, honekin irakaskuntzan laguntzeko sistema oso bat lortzeko helburutan.
- Galdera sortzailearen bukaerako ebaluazio sakonago bat burutzea.
- Termino erauzlearen ebaluazioa errepikatzea teknologia corpuseko zarata kendu ostean.
- Esaldien sinplifikazioa azterketa zein inplementazio sakonago bat burutzea. Honek galderen gramatikaltasun zein kopurua hobetuko lukeela uste da.
- Galderak sortzeko orduan dependentzien erabilera hobeago bat burutzea, kasuetan oinarritutako hautagai aukeraketa batetik dependentzietan oinarrituriko batera pasatuz.
- Galderak sortzeko transformazio erregelak idazteko patroi sistema bat lantzea. Tregex zein Tsurgeon patroiak interesgarriak izan litezke honetarako.
- Eduki aukeraketa burutzeko kontzeptu mapek izan dezaketen onura aztertzea.
- Termino erauzlearen inguruan buruturiko ebaluazioa beste corpus batekin errepikatzea, teknologia corpusean emaitzetan desbideraketa bat sortu izan dezakeen zarata bat detektatu baitzen.
- Esaldietan oinarrituriko galderak sortzetik diskurtso markatzaileetan oinarrituriko galdera sorkuntzara aurrera pauso bat ematea.
- Galderen kalitatea hobetzeko datu meatzaritzan oinarrituriko teknologietan pausoak ematea.

6.2.1 Bi moduluen integrazioan lehen pausoak

Jarraian etorkizuneko lanetatik erronka handiena suposatuko duenaren inguruko pauso batzuk aurkezten dira. Erronka bi moduluen arteko integrazio egoki bat lortzea izango da, irakaskuntza ingurune batean laguntzeko gai den tresna bat sortzeari begira. Integrazio honen aitzindari izan daitezkeen pausu batzuk eman dira proiektuaren azken egunetan. Integrazioa ahalik eta modu sinpleenean garatu da, baina etorkizunean burutu beharko den lanaren inguruko ideia bat sortzeko lagundu digu.

Integrazio posible bezala, galdera sortzailearen moduluaren irteera hartu da eta bertan agertzen diren terminoak detektatzen dituen sistema bat sortu da. Termino hauek pisatu eta galdera bakoitzari, bai galderan, bai erantzunean dauden terminoen pisuen batura eman zaio pisu gisa. Pisu hauekin galderaren testuarekiko garrantzia modelatzeko saiakera bat burutu da.

Sistema berri honen inguruan azaleko ebaluazio bat burutu da. Etorkizun batean ebaluazio hau modu sakonago batean burutuko den arren, sistemaren portaeraren ideia bat lortzeko aukera ematen du.

6.2.1.1 Erabilitako testuak

Ebaluazioa burutzeko erabili diren testuak zientzia eta teknologia arloko lau testu izan dira. Testu hauek QG-Maltiren ebaluaziorako erabilitako berdinak dira. DBH mailakotzat sailkatuak izan dira aditu batez. Hauen gaiak planeta, kontinentea, saguzarra eta artikoa dira. Testu guztiek luzera antzekoa dute (35 esaldi inguru), guztira 176 esaldi dituztelarik. Esaldien batazbesteko luzera 13 hitzekoa.

6.2.1.2 Ebaluazioaren metodologia

Ebaluazioan 2 esperimentu ezberdin burutu dira. Hauek derrigorrezko bigarren hezkuntzako euskarako irakasle baten laguntzaz burutu dira. Ebaluazioa burutzeko Seneko plataforma erabili da. Honekin bateragarria den formatuan sortu dira galderak eta bertara igo dira irakasleak ebaluatu ditzan.

6.2.1.3 Termino erauzlearen eraginkortasuna

Lehen esperimentuaren helburua galdera esanguratsuak aukeratzeko orduan termino erauzlearen eraginkortasuna neurtzea da. Honetarako 2 testu (*arti-*

koa eta *saguzar*) eman zaizkio irakasleari eta hauek ulertzeko interesgarriak izan daitezkeen 10 galdera sortzeko eskatu zaio bakoitzean.

6.2.1.4 Galderen esangura

Bigarren esperimentuaren helburua sistemak sortzen dituen galderek testuaren ulerkuntzan eman dezaketen laguntza neurtzea da. Honetarako irakasleari bi testu(*planeta* eta *kontinente*) eman zaizkio, eta sistemak hauen gainean sortu dituen hamarna galdera. Galdera bakoitzak testuaren ulermenean izan dezakeen garrantzia neurtzeko eskatu zaio 0tik 3rako zenbakiekin.

6.2.1.5 Lortutako emaitzak

Jarraian ebaluazioan burututako esperimentuen emaitzak aurkezten dira.

Termino erauzlearen eraginkortasuna
Lehen esperimentuak termino erauzleak galderen garrantzia neurtzeko orduan izan dezakeen eraginkortasuna neurtzea du helburu. Honetarako ebaluatzaileari bi testu eman zaizkio eta hauen ulermena neurtzeko hamarna galdera sortzeko eskatu zaio. Galdera hauetan terminoak bilatu dira termino erauzlearen bitartez eta bakoitzari pisu bat eman zaio.

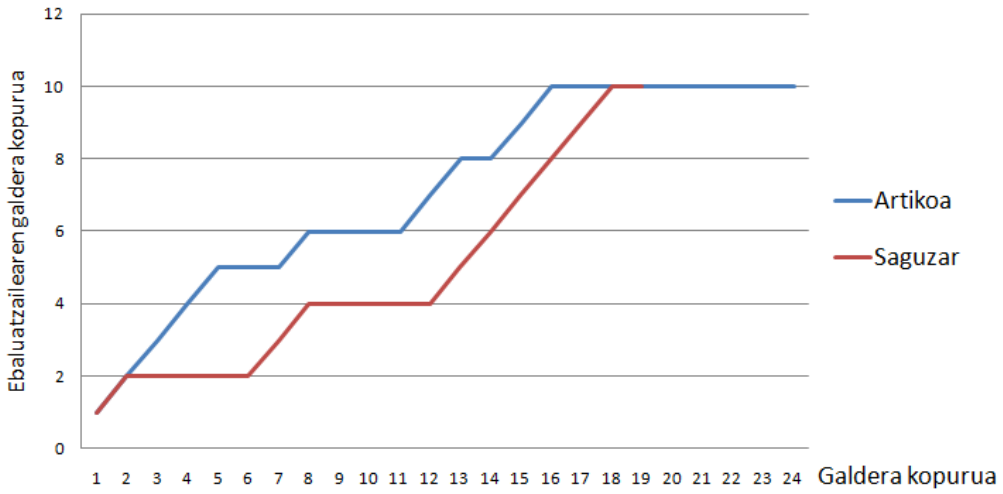
	Galderako termino kopurua		**Galderaren pisua**	
	Artikoa	**Saguzar**	**Artikoa**	**Saguzar**
Minimoa	2	0	0,0004	0
Batazbestekoa	2,6	1,5	0,0054286	0,0026368
Maximoa	5	4	0,035496	0,014122

Taula 6.1: Galderen termino kopurua eta pisua

Artikoa testuan galderako termino kopurua 2.6 da batazbestean. *Saguzar* testuan, aldiz, galderako termino kopurua 1.5ekoa da batazbestean. Galderen pisuan ere jaitsiera nabaria da, 0,0054286tik *Artikoa*n 0,0026368ra *Saguzar*rean. Jaitsiera hau *Saguzar* testuaren espezializazio mailaren ondorioa izan daiteke. Bertako terminoak zientzia eta teknologiako corpusean agertzeko joera txikiagoa dute *Artikoa* testuko terminoek baino.

Irakasleak sorturiko galderak eta sistemak sorturiko galderak multzo berean jarri eta pisuaren arabera ordenatu dira. Pisatze sistema egokia bada,

irakasleak sorturiko galderak aurrealdean egon beharko lukete, hauek baitira benetan sortu nahi diren galderak. Zerrenda honen estatistikak 6.1 irudian ikus daitezke.



Irudia 6.1: Ebaluatzailearen zein sistemaren esaldiak ordenaturik

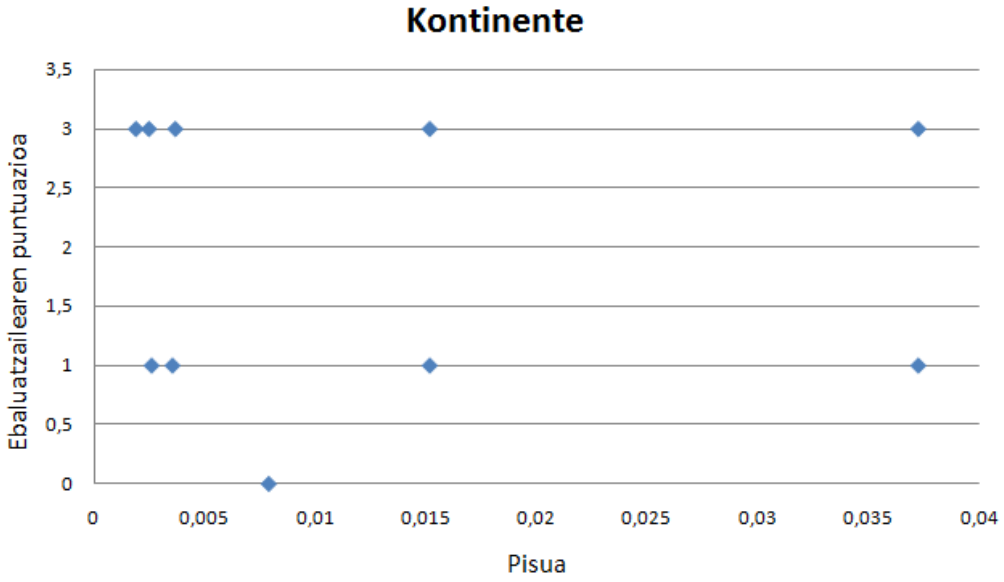
Bertan, zerrendan agertzen diren lehen x galderetatik zenbat diren ebaluatzaileak sorturikoak adierazten da. Irudian ebaluatzailearen galderak hasiera aldean kokatzen direla ikus daiteke, bereziki *Artikoa* testuan. Honek ebaluatzailearen galderei pisu handienak ematen zaizkiela esan nahi du, termino erauzleak eduki aukeraketa burutzeko orduan izan dezakeen eraginkortasuna frogatuz.

Galderaren esangura

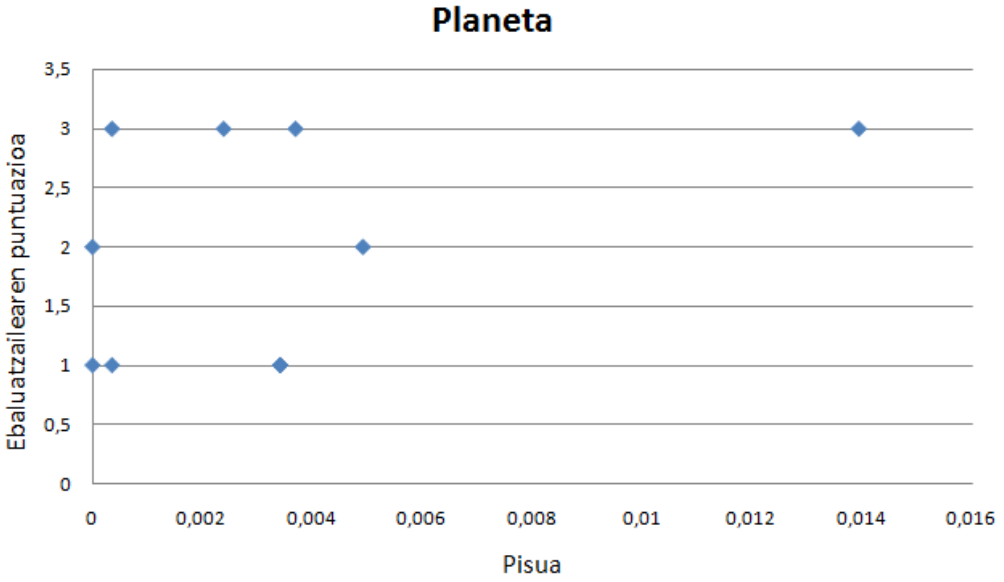
Bigarren esperimentuan sistemak sortzen dituen galderen esangura maila neurtzeaz gain galderen pisua eta ebaluatzaileak emandako pisuaren artean korrelaziorik dagoen neurtuko da. Honetarako ebaluatzaileari bi testu(*Kontinente* eta *Planeta*) eman zaizkio eta testuko hamarna galdera. Galdera bakoitza testuarekiko duen esanguraren arabera 0tik 3ra puntuatzeko eskatu zaio. *Kontinente* testuan galderek duten puntuazioan bataz bestekoa 2 da eta *Planeta* testuan, aldiz, 1.9. Bi batazbestekoek galderak testuarekiko esangura altua dutela ondorioztatzera eramaten gaituzte.

Datuak ikusirik irakaslearen puntuazioen eta galderen pisuaren artean korrelaziorik dagoen neurtu da. Honetarako bi testuen sakabanaketa grafikoak

zein korrelazio koefizienteak kalkulatu dira. Sakabanaketa grafikoak 6.2 eta 6.3 irudietan ikus daitezke.



Irudia 6.2: Sakabanaketa grafikoa: Kontinente



Irudia 6.3: Sakabanaketa grafikoa: Planeta

Sakabanaketa grafikoetan ez dagoela ez ordenazio ez patroi berezirik ikus daiteke. Hau korrelazio koefizientean ere ikus daiteke. *Planeta* testurako korrelazio koefizientea 0.37 da eta *Kontinente* testukoa -0,01. Lehenengoan pisua igo ahala irakaslearen puntuazioa igotzen dela adierazten da, baina bigarrenean bi aldagaien arteko erlaziorik ez dagoela ikus da.

Honek datuak hurbilagotik ikustera eraman gaitu eta esaldi berdinetik sorturiko zenbait galderak irakaslearen partetik puntuazio zeharo ezberdina jasotzen zutela ikusi dugu. Hona hemen adibide bat:

-Zer bete zuten konbekzio-korrontea eratutakoan , hiru milimetroko bolatxoz , bolatxoen eragina ikusteko?
-hondoaren azalera

-Noiz bete zuten hondoaren azalera erdia hiru milimetroko bolatxoz , bolatxoen eragina ikusteko?
-konbekzio-korrontea eratutakoan

Bi galdera horien jatorrizko esaldia berbera denez sistemak pisu bera ematen die. Aldiz, irakasleak *konbekzio-korrontea* erantzunean duen galderari 3ko puntuazioa eman dio eta besteari 1ekoa, *konbekzio-korronteek* testuarekiko zeresan handiagoa baitzuten. Honek galderen pisua neurtzeko erabilitako metodoa baino zerbait gehiago behar dela ondorioztatzera eramaten gaitu. Terminoek galderaren esanguran zeresana badutela dirudi, baina galderen pisua neurtzeko metodoa hobetzeko dago oraindik.

Bestalde, irakaslearekin izandako elkarrizketetan galderak hobetzeko esaldi mailatik goragoko teknikaren bat erabiltzeko beharra ikusi da etorkizunerako. Berak zioenez, sistemak sorturiko galderak testuaren ulermena lantzeko galderak baino gehiago memoria lantzeko galderak ziren. Izan ere, esaldi mailan galderak sortu izanak esaldi bateko datu espezifikoen gaineko galderak sortzera mugatzen gaitu eta ez irakasleak zioen moduan testuan zeuden ideien arteko erlazioen inguruko galderak sortzera. Nolabait esanda, testuan ager daitezken arrazonamendu, ondorio eta antzekoen inguruan galdetzea egokiagoa izango litzateke. Honek galderen eraikuntzan beste aurrera pausu bat eman beharko dela ondorioztatzera eraman gaitu testuaren ezagutza errepresentatzeko sistemaren bat baliatuz.

Ikusi den moduan, termino erauzle eta galdera sortzaile moduluena arteko integrazioa erronka bat izango da etorkizunerako. Ziurrenik burutu dugun

terminoen batura baino teknika sofistikatuagoak erabiltzea beharrezkoa izango da galderak sailkatzeko eta galdera eraikuntzan esaldi mailatik haratagoko teknikak aztertu beharko dira, baina aurre ebaluazio honetan ikusiriko zenbait datuk bide egokia jorratzen ari garen irudipena uzten digute.

I Eranskina

Atazen denbora estimazioa

Gaien Aurkibidea

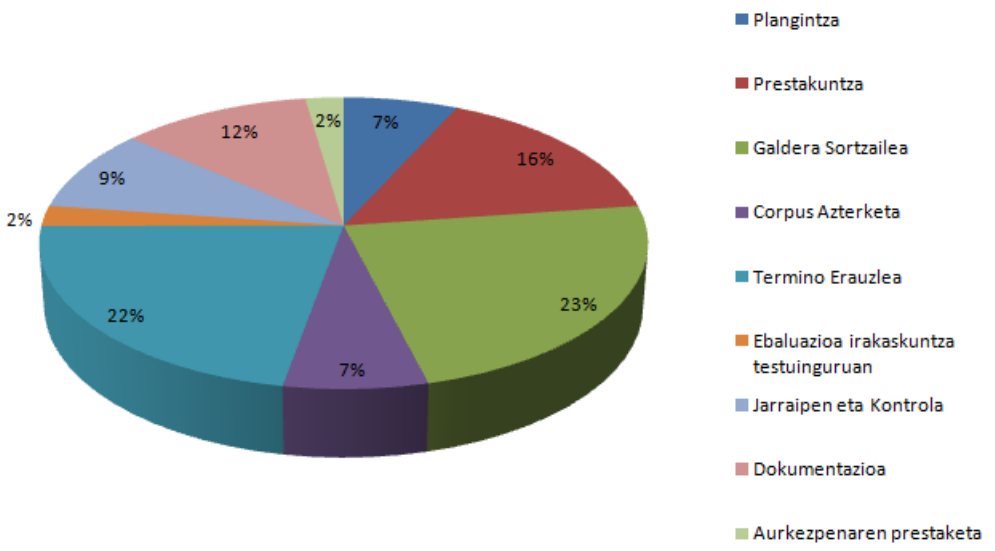
I.a Denbora estimazioak

Jarraian ataza eta azpiataza bakoitzari esleitutako denbora estimazioak aurkezten dira.

Ataza	**Estimatutako Orduak**
Plangintza	30
Prestakuntza	70
Galdera Sortzailea	102
Analisia	20
Diseinua	12
Inplementazioa	33
Probak	21
Ebaluazioa	16
Corpus Azterketa	30
Analisia	7
Eskurapena	16
Konparaketa	4
Aukeraketa	3
Termino erauzlea	97
Analisia	19
Diseinua	12
Inplementazioa	35
Probak	14
Ebaluazioa	17
Ebaluazioa irakaskuntza testuinguruan	10
Jarraipen eta kontrola	40
Dokumentazioa	50
Aurkezpen prestaketa	10
Guztira	**439**

Taula I.1: Atazen denbora estimazioa

I.b Datuak laburtzen



Irudia I.1: Denboren Estimazioa

Denbora estimazioen irudiari erreparatzen badiogu, proiektuan denbora gehien beharko diren atazak garapenera dedikaturiko moduluak direla ikusten dugu, galdera sortzaile modulua eta termino erauzlea hain zuen ere.

Hauen atzetik prestakuntza ere nabarmentzen da bere lan kargagatik, proiektua ikerkuntzara orientatua izanik denbora asko dedikatuko baitzaio artikulu irakurketa zein aurrekarien azterketari.

Azkenik, plangintza, jarraipen eta dokumentazioari dagokin lan karga ere ezin da ahaztu, gradu amaierako proiektu bat izanik ataza honek ere garrantzia dezentekoa hartzen baitu.

II Eranskina

Termino erauzleen ebaluazioaren datu zehatzagoak

Gaien Aurkibidea

II.a Ebaluatzaileen adostasuna

II.a.1 Historia (Hitz bakarrekoak)

	2. Ebaluatzailea			
1. Ebaluatzailea		Bai	Ez	
	Bai	45	14	59
	Ez	15	25	40
		60	39	99

Taula II.1: Konkordantzia taula: TE1, historia, hitz bakarrekoak

Kappa: 0,3892788768
Adostasuna: 0,7070707071

	2. Ebaluatzailea			
1. Ebaluatzailea		Bai	Ez	
	Bai	54		58
	Ez	16	26	42
		70	30	100

Taula II.2: Konkordantzia taula: TE3, historia, hitz bakarrekoak

Kappa: 0,5726495726
Adostasuna: 0,8

	2. Ebaluatzailea			
1. Ebaluatzailea		Bai	Ez	
	Bai	65	3	68
	Ez	22	10	32
		87	13	100

Taula II.3: Konkordantzia taula: Erauzterm, historia, hitz bakarrekoak

Kappa: 0,3184296619
Adostasuna: 0,75

II.a.2 Historia (Hitz anitzekoak)

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	30	13	43
	Ez	6	50	56
		36	63	99

Taula II.4: Konkordantzia taula: TE1, historia, hitz anitzekoak

Kappa: 0,6019047619
Adostasuna: 0,8080808081

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	89	3	92
	Ez	3	5	8
		82	8	100

Taula II.5: Konkordantzia taula: TE3, historia, hitz anitzekoak

Kappa: 0,5923913043
Adostasuna: 0,94

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	72	3	75
	Ez	13	12	25
		85	15	100

Taula II.6: Konkordantzia taula: Erauzterm, historia, hitz anitzekoak

Kappa: 0,5076923077
Adostasuna: 0,84

II.a.3 Teknologia Corpusa (Hitz bakarrekoak)

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	22	11	33
	Ez	16	51	77
		38	62	100

Taula II.7: Konkordantzia taula: TE1, teknologia, hitz bakarrekoak

Kappa: 0,4120209059
Adostasuna: 0,73

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	2	2	4
	Ez	6	90	96
		8	92	100

Taula II.8: Konkordantzia taula: TE3, teknologia, hitz bakarrekoak

Kappa: 0,2957746479
Adostasuna: 0,92

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	15	7	22
	Ez	5	73	78
		20	80	100

Taula II.9: Konkordantzia taula: Erauzterm, teknologia, hitz bakarrekoak

Kappa: 0,6385542169
Adostasuna: 0,88

II.a.4 Teknologia Corpusa (Hitz anitzekoak)

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	31	9	40
	Ez	10	49	59
		41	58	99

Taula II.10: Konkordantzia taula: TE1, teknologia, hitz anitzekoak

Kappa: 0,6030808187
Adostasuna: 0,8080808081

	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	23	5	28
	Ez	8	64	72
		31	69	100

Taula II.11: Konkordantzia taula: TE3, teknologia, hitz anitzekoak

Kappa: 0,6878001921
Adostasuna: 0,87

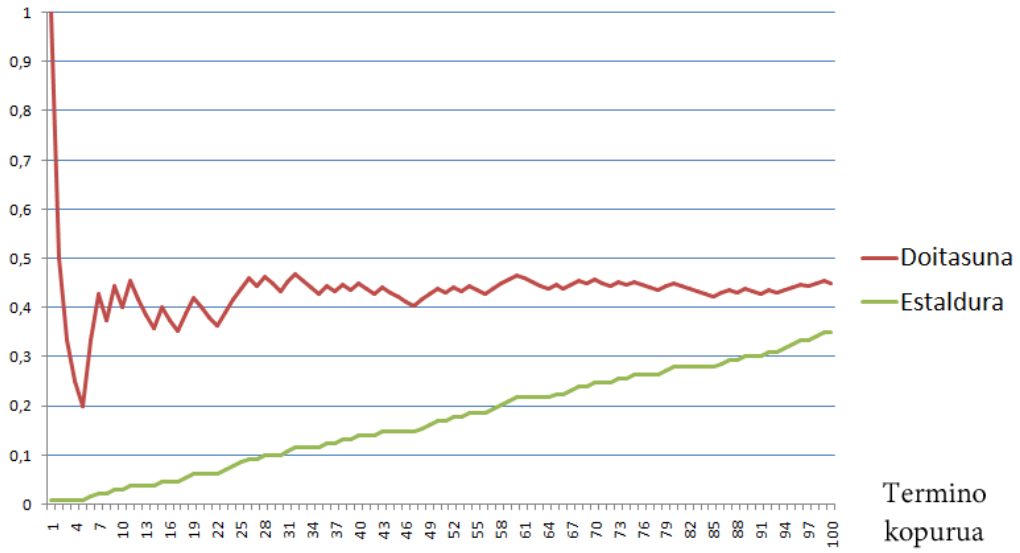
	2. Ebaluatzailea			
		Bai	Ez	
1. Ebaluatzailea	Bai	28	3	31
	Ez	12	57	69
		40	60	100

Taula II.12: Konkordantzia taula: Erauzterm, teknologia, hitz anitzekoak

Kappa: 0,6753246753
Adostasuna: 0,85

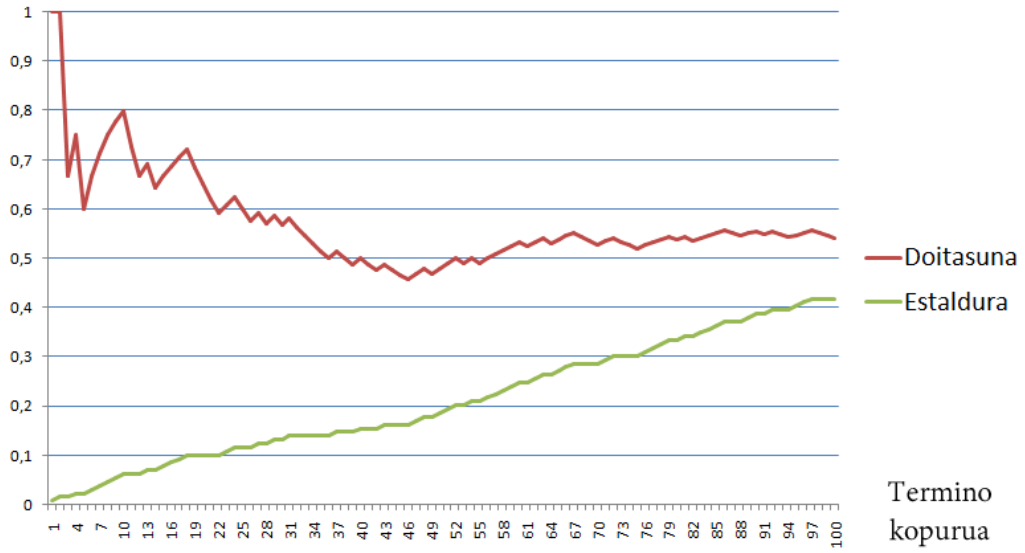
II.b Sistemen ebaluazioa

II.b.1 Historia (Hitz bakarrekoak)



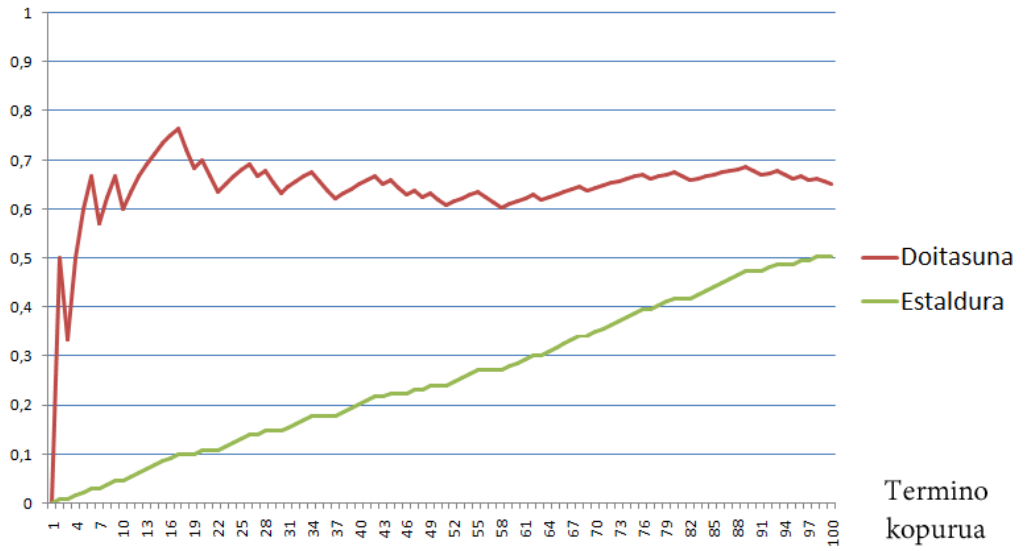
Irudia II.1: Ebaluazio grafikoa: TE1, historia, hitz bakarrekoak

Doitasuna: 0,45
Estaldura: 0,348837209



Irudia II.2: Ebaluazio grafikoa: TE2, historia, hitz bakarrekoak

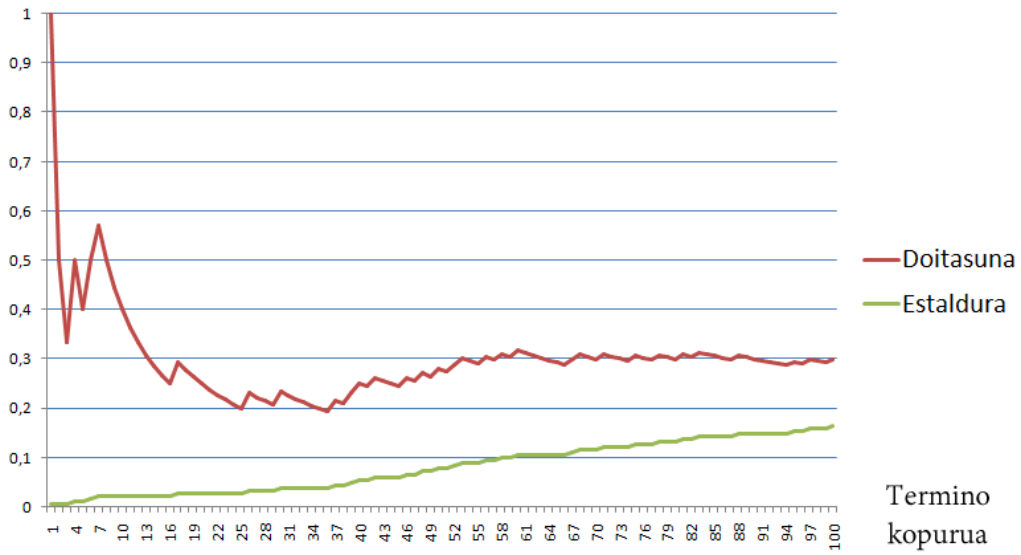
Doitasuna: 0,54
Estaldura: 0,418604651



Irudia II.3: Ebaluazio grafikoa: Erauzterm, historia, hitz bakarrekoak

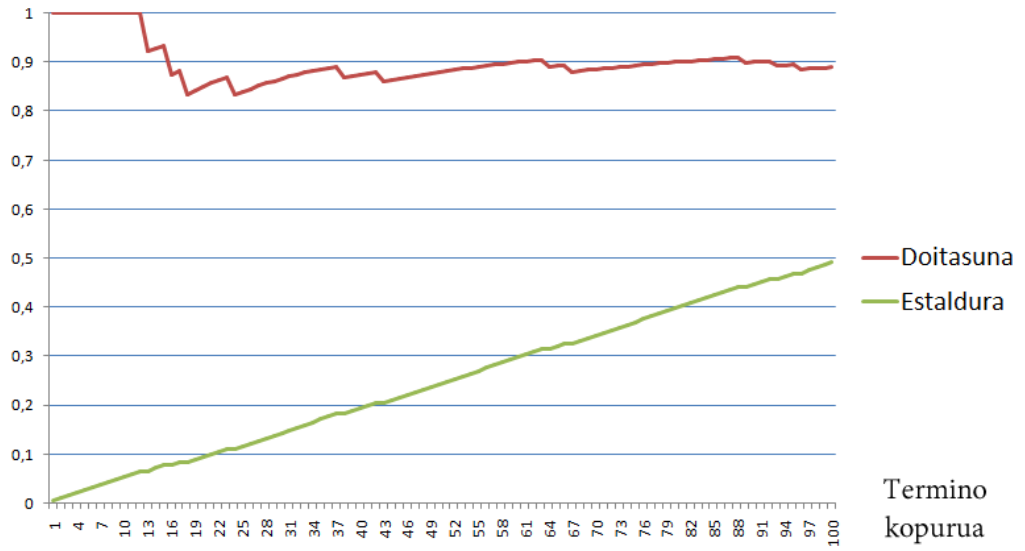
Doitasuna: 0,65
Estaldura: 0,503875969

II.b.2 Historia (Hitz anitzekoak)



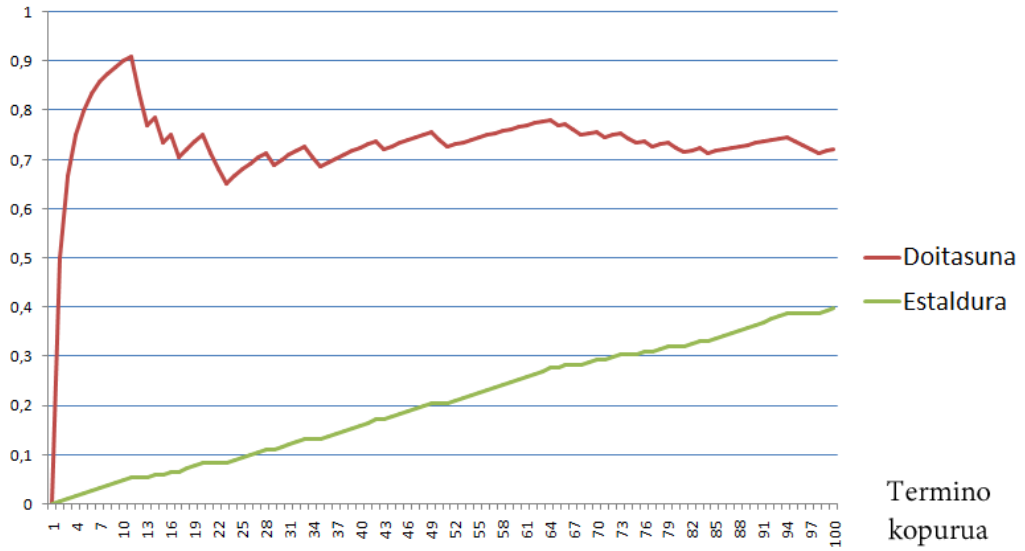
Irudia II.4: Ebaluazio grafikoa: TE1, historia, hitz anitzekoak

Doitasuna: 0,3
Estaldura: 0,165745856



Irudia II.5: Ebaluazio grafikoa: TE2, historia, hitz anitzekoak

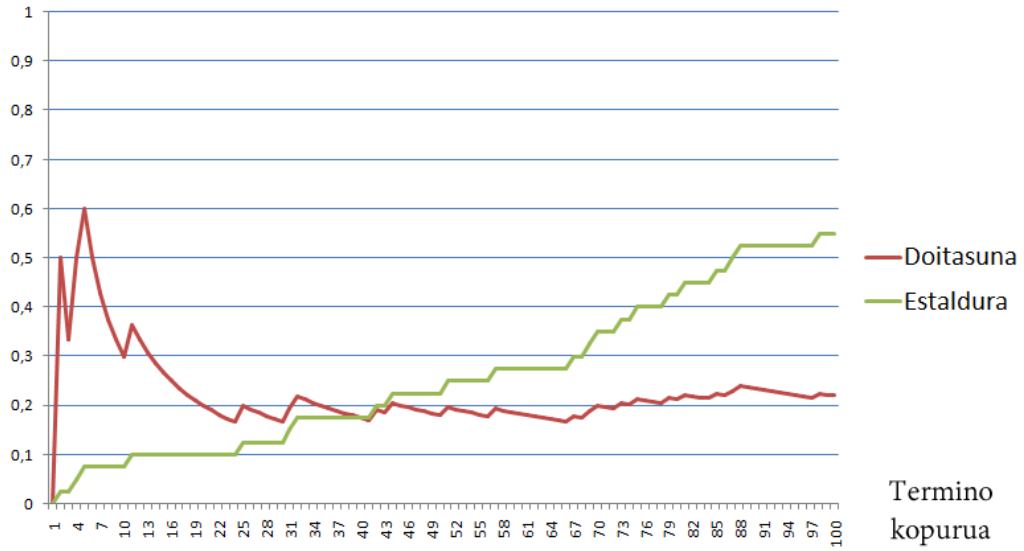
Doitasuna: 0,89
Estaldura: 0,491712707



Irudia II.6: Ebaluazio grafikoa: Erauzterm, historia, hitz anitzekoak

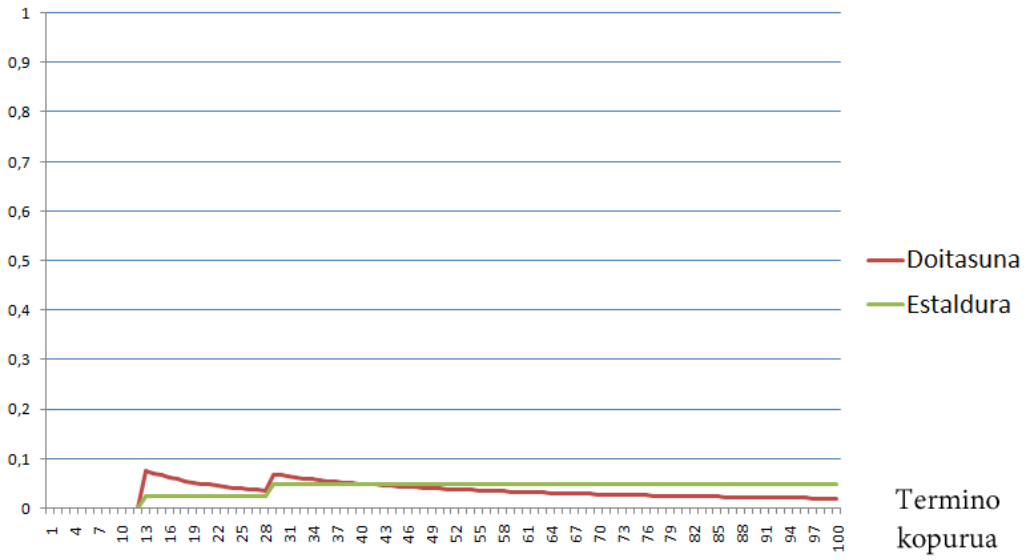
Doitasuna: 0,72
Estaldura: 0,397790055

II.b.3 Teknologia (Hitz bakarrekoak)



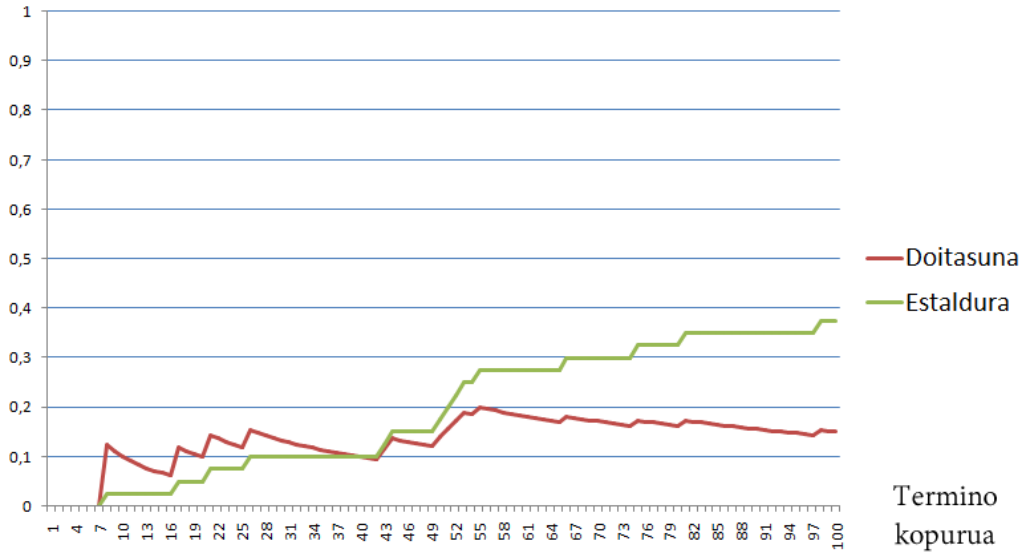
Irudia II.7: Ebaluazio grafikoa: TE1, teknologia, hitz bakarrekoak

Doitasuna: 0,22
Estaldura: 0,55



Irudia II.8: Ebaluazio grafikoa: TE2, teknologia, hitz bakarrekoak

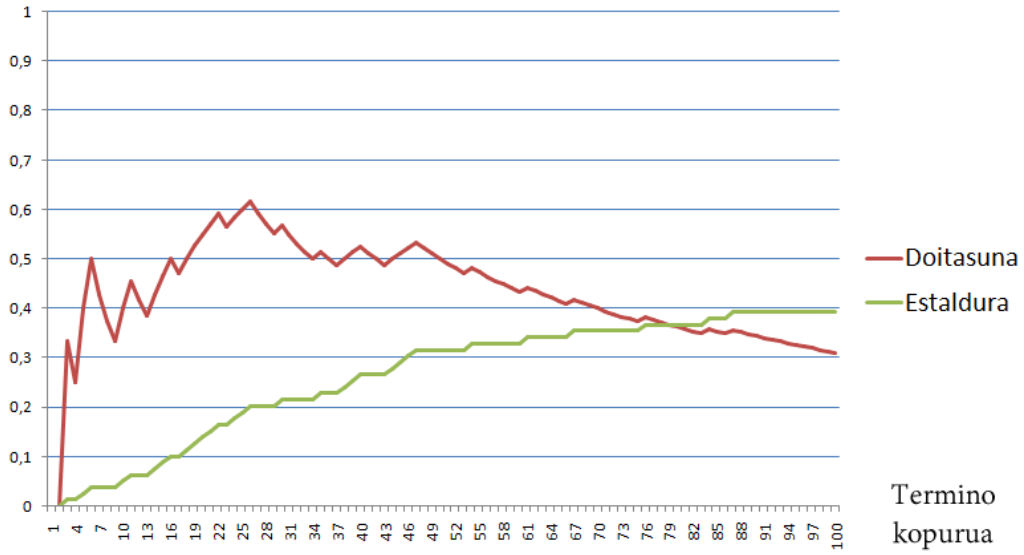
Doitasuna: 0,02
Estaldura: 0,05



Irudia II.9: Ebaluazio grafikoa: Eraузterm, teknologia, hitz bakarrekoak

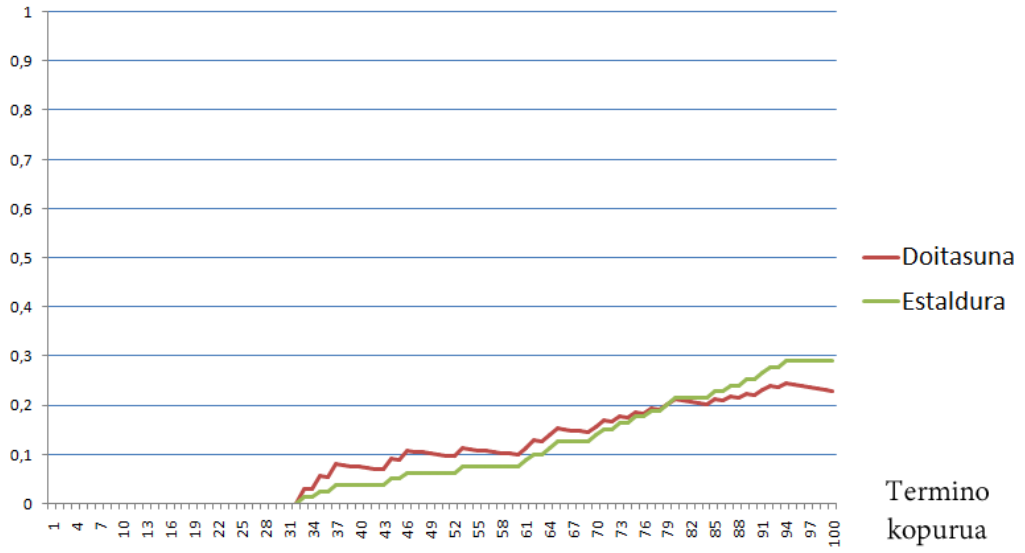
Doitasuna: 0,15
Estaldura: 0,375

II.b.4 Teknologia (Hitz anitzekoak)



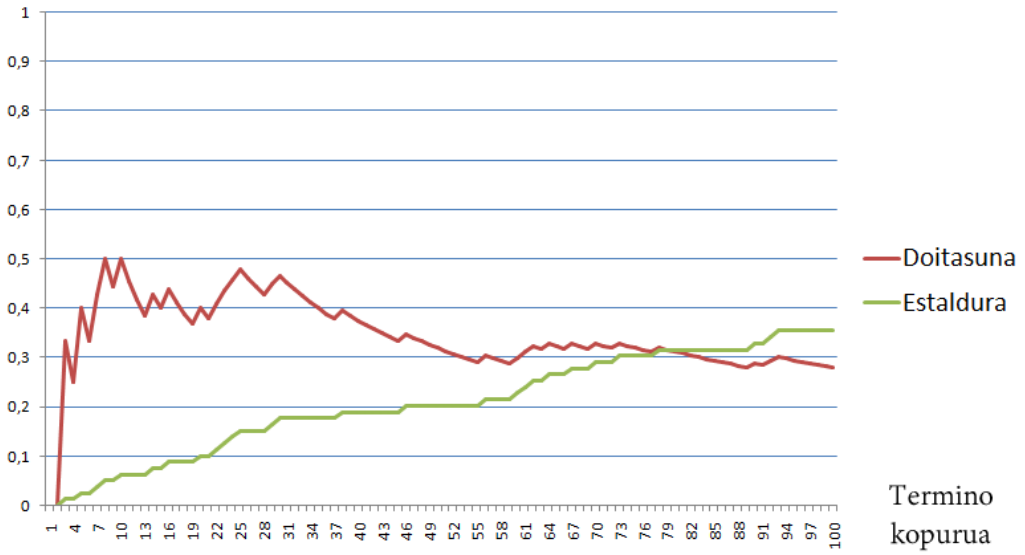
Irudia II.10: Ebaluazio grafikoa: TE1, teknologia, hitz anitzekoak

Doitasuna: 0,31
Estaldura: 0,392405063



Irudia II.11: Ebaluazio grafikoa: TE2, teknologia, hitz anitzekoak

Doitasuna: 0,23
Estaldura: 0,291139241



Irudia II.12: Ebaluazio grafikoa: Erauzterm, teknologia, hitz anitzekoak

Doitasuna: 0,28
Estaldura: 0,35443038

III Eranskina

Detektatutako terminoen adibideak

Gaien Aurkibidea

Jarraian *Kontinente* testuan detektatu diren termino esanguratsuak erakusten dira. Hau ikusteko lehen 500, 1000 eta 3000 terminoekin adibideak aurkezten dira

III.a Lehen 500 terminoak erabilita

Kontinenteen jitoa imitatzeko , bolatxoak . Orain dela 300 milioi urte , Pangea superkontinentea zegoen ; orain dela 200 milioi urte inguru , bereizten hasi zen , eta gaur ezagutzen ditugun kontinenteak sortu ziren . Hemendik milioi urte batzuetara , berriz elkartuko dira seguru asko ... Lurreko kontinenteak etengabe ari dira mugitzen , elkartzen eta urruntzen , lurrazalaren azpian dagoen mantu likidoaren konbekzio-korronteek bultzatuta , onartuen dagoen teoriari jarraitzen badiogu . Bada , New York Unibertsitateko fisikari batzuk saiatu dira azaltzen zergatik gertatzen den urruntze eta aldentze zikliko hori . Hain zuzen , lurrazalaren lodiera aldakorrak konbekzio-korronteetan duen eragina simulatu dute . Esperimentu erraz baten bidez (0,007634) egin zuten : hiru litroko tanga bat urez eta glizerinaz bete zuten , eta azpian xafla (0,014286) bero bat jarrita berotu zuten likidoa . Ura berotzearekin batera , konbekzio-korronte bat eratu zen : azpialdean berotutako likidoak (0,003279) gora egiten zuen dentsitatea galdutakoan , eta gaineko (0,002203) likido (0,003279) hotzagoak beherantz , dentsitatea handiagoa zuelako . Hala , korronte zirkular bat eratu zen . Konbekzio-korrontea eratutakoan , hondoaren (0,004237) azalera erdia (0,029412) hiru milimetroko bolatxoz bete zuten , bolatxoen eragina ikusteko . Eta , ikusi zutenez , eragina handia izan zen : askoz maizago aldarazten zuten korrontearen noranzkoa , eta , aldi berean , haiek ere tangaren alde batetik bestera mugitzen ziren . Sartu zituztenean , bolatxoak tangaren alde (0,006623) batean pilatu ziren , likidoa noranzko (0,008929) batean zebilelako biraka . Hala , likidoari (0,003279) beroa ematen zion xaflaren (0,014286) eta likidoaren artean jarri ziren bolatxoak , eta beroari bidea (0,007634) oztopatu zioten . Bolatxorik ez zuen aldean (0,006623) , berriz , hasierako (0,010309) indarraz berotzen jarraitu zuen urak . Beraz , alde (0,006623) horretako ura gora egiten hasi zen ; hau da , konbekzio-korronteak bira eman zuen . Korronteak bira emandakoan , bolatxoei bultzada beste aldetik (0,006623) iristen hasi zitzaien , eta mugitzen hasi ziren , beste muturreraino iritsi ziren arte . Orduan , berriz hasi zen prozesua . Bolatxoak eta kontinenteak , ' gauza bera ' . Zer zerikusi du - galdetuko zion honezkero bat baino gehiago bere buruari-fisikari horien esperimentuak Lurreko kontinenteen mugimenduarekin (0,006494) ? Bada , fisikariek eurek diote sistemak (0,002041) oso antzekoak direla . Lurraren (0,002915) kasuan bolatxoak kontinenteak lirateke , eta , tangaren hondoan

(0,004237) egon beharrean , gainean (0,002203) flotatzen egongo lirateke . Likidoa lurrazalaren azpian dagoen mantua litzateke . Une honetan , Lurraren (0,002915) nukleoak (0,005263) igortzen duen beroaren eraginez , mantuan konbekzio-korronteak noranzko (0,008929) batean biratzen ari dira , eta korronte horiek bultzatuta ari dira mugitzen kontinenteak . Uneren batean , ordea , Lurraren (0,002915) eremuren (0,00369) batean bat egingo dute kontinente guztiek , hau da , garai (0,002183) batean izan zen Pangearen itxurako (0,013158) superkontinente bat eratuko da . Kontinenteen azpiko lurrazala askoz lodiagoa denez ozeanoen azpian dagoena (0,003021) baino , superkontinenteak bolatxoen antzeko funtzioa izango duela esan dute ikertzaileek (0,002227) . Antza denez , Lurraren (0,002915) mantuak igortzen duen beroa superkontinentearen azpian metatuko da , eta itsasoen azpiko mantua kontinentearen azpikoa baino hotzago egotera pasatuko da . Horren eraginez , hotzago dagoen materiala hondoratuz joango da , eta kontinenteak gune (0,003663) batean pilatu dituen konbekzio-korronteek bira emango dute . Orduan , kontinenteak elkartzea ez , baizik elkarrengandik bereiztea eragingo dute korronte sortu berriek . Zientzialarien esanean , hainbatean gertatu da hori Lurraren (0,002915) historian ; 200-250 milioi urtean behin , hain zuzen . Hortaz , Pangea baino lehenago beste elkartze eta bereizte batzuk gertatu ziren , nonbait , eta hemendik aurrera ere gertatuko omen dira . (OR) : Plaka tektoniko deritzen egituren (0,00495) zati (0,016393) dira gaur egun ezagutzen ditugun kontinenteak . Plaka tektoniko horiek mantuaren gainean (0,002203) flotatzen ari dira . Bolatxoak : New York Unibertsitateko taldeak egin zuen esperimentuaren eskema . Bolatxoak tangaren alde (0,006623) batean pilatzean , konbenkzio-korronteak bira ematen du .

III.b Lehen 1000 terminoak erabilita

Kontinenteen jitoa imitatzeko , bolatxoak . Orain dela 300 milioi urte , Pangea superkontinentea zegoen ; orain dela 200 milioi urte inguru , bereizten hasi zen , eta gaur ezagutzen ditugun kontinenteak sortu ziren . Hemendik milioi urte batzuetara , berriz elkartuko dira seguru asko ... Lurreko kontinenteak etengabe ari dira mugitzen , elkartzen eta urruntzen , lurrazalaren azpian dagoen mantu likidoaren konbekzio-korronteek bultzatuta , onartuen dagoen teoriari (0,00111) jarraitzen badiogu . Bada , New York Unibertsitateko fisikari batzuk saiatu dira azaltzen zergatik gertatzen den urruntze eta aldentze zikliko hori . Hain zuzen , lurrazalaren lodiera aldakorrak konbekzio-korronteetan duen eragina simulatu dute . Esperimentu (0,001263) erraz baten bidez (0,007634) egin zuten : hiru litroko tanga bat urez (0,001595) eta glizerinaz bete zuten , eta azpian xafla (0,014286) bero bat jarrita berotu zuten likidoa . Ura (0,001595) berotzearekin batera , konbekzio-korronte bat eratu zen : azpialdean berotutako likidoak (0,003279) gora egiten zuen dentsitatea galdutakoan , eta gaineko (0,002203) likido (0,003279) hotzagoak beherantz , dentsitatea handiagoa zuelako . Hala , korronte (0,001642) zirkular bat eratu zen . Konbekzio-korrontea eratutakoan , hondoaren (0,004237) azalera erdia (0,029412) hiru milimetroko (0,001721) bolatxoz bete zuten , bolatxoen eragina ikusteko . Eta , ikusi zutenez , eragina handia izan zen : askoz maizago aldarazten zuten korrontearen (0,001642) noranzkoa , eta , aldi berean , haiek ere tangaren alde batetik bestera mugitzen ziren . Sartu zituztenean , bolatxoak tangaren alde (0,006623) batean pilatu ziren , likidoa noranzko (0,008929) batean zebilelako biraka . Hala , likidoari (0,003279) beroa ematen zion xaflaren (0,014286) eta likidoaren artean jarri ziren bolatxoak , eta beroari bidea (0,007634) oztopatu zioten . Bolatxorik ez zuen aldean (0,006623) , berriz , hasierako (0,010309) indarraz berotzen jarraitu zuen urak (0,001595) . Beraz , alde (0,006623) horretako ura (0,001595) gora egiten hasi zen ; hau da , konbekzio-korronteak bira eman zuen . Korronteak (0,001642) bira emandakoan , bolatxoei bultzada beste aldetik (0,006623) iristen hasi zitzaien , eta mugitzen hasi ziren , beste muturreraino iritsi ziren arte . Orduan , berriz hasi zen prozesua (0,001399) . Bolatxoak eta kontinenteak , ' gauza bera ' . Zer zerikusi du - galdetuko zion honezkero bat baino gehiago bere buruari-fisikari horien esperimentuak (0,001263) Lurreko kontinenteen mugimenduarekin (0,006494) ? Bada , fisikariek eurek diote sistemak (0,002041) oso antzekoak direla . Lurraren (0,002915) kasuan bolatxoak kontinenteak lirateke , eta , tangaren hondoan (0,004237) egon beharrean , gainean (0,002203) flotatzen egongo lirateke . Likidoa lurrazalaren azpian dagoen mantua litzateke . Une honetan , Lurraren (0,002915) nukleoak (0,005263) igortzen duen beroaren eraginez , mantuan

konbekzio-korronteak noranzko (0,008929) batean biratzen ari dira , eta korronte (0,001642) horiek bultzatuta ari dira mugitzen kontinenteak . Uneren batean , ordea , Lurraren (0,002915) eremuren (0,00369) batean bat egingo dute kontinente guztiek , hau da , garai (0,002183) batean izan zen Pangearen itxurako (0,013158) superkontinente bat eratuko da . Kontinenteen azpiko lurrazala askoz lodiagoa denez ozeanoen azpian dagoena (0,003021) baino , superkontinenteak bolatxoen antzeko funtzioa izango duela esan dute ikertzaileek (0,002227) . Antza denez , Lurraren (0,002915) mantuak igortzen duen beroa superkontinentearen azpian metatuko da , eta itsasoen azpiko mantua kontinentearen azpikoa baino hotzago egotera pasatuko da . Horren eraginez , hotzago dagoen materiala hondoratuz joango da , eta kontinenteak gune (0,003663) batean pilatu dituen konbekzio-korronteek bira emango dute . Orduan , kontinenteak elkartzea ez , baizik elkarrengandik bereiztea eragingo dute korronte (0,001642) sortu berriek . Zientzialarien (0,001261) esanean , hainbatean gertatu da hori Lurraren (0,002915) historian ; 200-250 milioi urtean behin , hain zuzen . Hortaz , Pangea baino lehenago beste elkartze eta bereizte batzuk gertatu ziren , nonbait , eta hemendik aurrera ere gertatuko omen dira . (OR) : Plaka tektoniko deritzen egituren (0,00495) zati (0,016393) dira gaur egun ezagutzen ditugun kontinenteak . Plaka tektoniko horiek mantuaren gainean (0,002203) flotatzen ari dira . Bolatxoak : New York Unibertsitateko taldeak (0,001692) egin zuen esperimentuaren (0,001263) eskema . Bolatxoak tangaren alde (0,006623) batean pilatzean , konbenkzio-korronteak bira ematen du .

III.c Lehen 3000 terminoak erabilita

Kontinenteen jitoa imitatzeko , bolatxoak . Orain dela 300 milioi urte (0,000505) , Pangea superkontinentea zegoen ; orain dela 200 milioi urte (0,000505) inguru , bereizten hasi zen , eta gaur ezagutzen ditugun kontinenteak sortu ziren . Hemendik milioi urte (0,000505) batzuetara , berriz elkartuko dira seguru asko ... Lurreko kontinenteak etengabe ari dira mugitzen , elkartzen eta urruntzen , lurrazalaren (0,000658) azpian (0,00089) dagoen mantu (0,00051) likidoaren konbekzio-korronteek bultzatuta , onartuen dagoen teoriari (0,00111) jarraitzen badiogu . Bada , New York Unibertsitateko fisikari batzuk saiatu dira azaltzen zergatik gertatzen den urruntze eta aldentze zikliko hori . Hain zuzen , lurrazalaren (0,000658) lodiera (0,000926) aldakorrak konbekzio-korronteetan duen eragina (0,000998) simulatu dute . Esperimentu (0,001263) erraz baten bidez (0,007634) egin zuten : hiru litroko tanga (0,000802) bat urez (0,001595) eta glizerinaz bete zuten , eta azpian (0,00089) xafla (0,014286) bero bat jarrita berotu zuten likidoa . Ura (0,001595) berotzearekin batera , konbekzio-korronte bat eratu zen : azpialdean berotutako likidoak (0,003279) gora egiten zuen dentsitatea galdutakoan , eta gaineko (0,002203) likido (0,003279) hotzagoak beherantz (0,000894) , dentsitatea handiagoa zuelako . Hala , korronte (0,001642) zirkular bat eratu zen . Konbekzio-korrontea eratutakoan , hondoaren (0,004237) azalera (0,000898) erdia (0,029412) hiru milimetroko (0,001721) bolatxoz bete zuten , bolatxoen eragina (0,000998) ikusteko . Eta , ikusi zutenez , eragina (0,000998) handia izan zen : askoz maizago aldarazten zuten korrontearen (0,001642) noranzkoa , eta , aldi berean , haiek ere tangaren (0,000802) alde batetik bestera mugitzen ziren . Sartu zituztenean , bolatxoak tangaren (0,000802) alde (0,006623) batean pilatu ziren , likidoa noranzko (0,008929) batean zebilelako biraka . Hala , likidoari (0,003279) beroa ematen zion xaflaren (0,014286) eta likidoaren artean jarri ziren bolatxoak , eta beroari bidea (0,007634) oztopatu zioten . Bolatxorik ez zuen aldean (0,006623) , berriz , hasierako (0,010309) indarraz (0,000745) berotzen jarraitu zuen urak (0,001595) . Beraz , alde (0,006623) horretako ura (0,001595) gora egiten hasi zen ; hau da , konbekzio-korronteak bira eman zuen . Korronteak (0,001642) bira emandakoan , bolatxoei bultzada beste aldetik (0,006623) iristen hasi zitzaien , eta mugitzen hasi ziren , beste muturreraino (0,000441) iritsi ziren arte . Orduan , berriz hasi zen prozesua (0,001399) . Bolatxoak eta kontinenteak , ' gauza bera ' . Zer zerikusi du - galdetuko zion honezkero bat baino gehiago bere buruari-fisikari horien esperimentuak (0,001263) Lurreko kontinenteen mugimenduarekin (0,006494) ? Bada , fisikariek eurek diote sistemak (0,002041) oso antzekoak direla . Lurraren (0,002915) kasuan (0,000456) bolatxoak kontinenteak lirateke , eta , tangaren (0,000802) hon-

doan (0,004237) egon beharrean (0,000409) , gainean (0,002203) flotatzen egongo lirateke . Likidoa lurrazalaren (0,000658) azpian (0,00089) dagoen mantua (0,00051) litzateke . Une (0,000627) honetan , Lurraren (0,002915) nukleoak (0,005263) igortzen duen beroaren eraginez (0,000998) , mantuan (0,00051) konbekzio-korronteak noranzko (0,008929) batean biratzen ari dira , eta korronte (0,001642) horiek bultzatuta ari dira mugitzen kontinenteak . Uneren (0,000627) batean , ordea , Lurraren (0,002915) eremuren (0,00369) batean bat egingo dute kontinente guztiek , hau da , garai (0,002183) batean izan zen Pangearen itxurako (0,013158) superkontinente bat eratuko da . Kontinenteen azpiko (0,00089) lurrazala (0,000658) askoz lodiagoa denez ozeanoen azpian (0,00089) dagoena (0,003021) baino , superkontinenteak bolatxoen antzeko funtzioa (0,00034) izango duela esan dute ikertzaileek (0,002227) . Antza denez , Lurraren (0,002915) mantuak (0,00051) igortzen duen beroa superkontinentearen azpian (0,00089) metatuko da , eta itsasoen azpiko (0,00089) mantua (0,00051) kontinentearen azpikoa (0,00089) baino hotzago egotera pasatuko da . Horren eraginez (0,000998) , hotzago dagoen materiala hondoratuz joango da , eta kontinenteak gune (0,003663) batean pilatu dituen konbekzio-korronteek bira emango dute . Orduan , kontinenteak elkartzea ez , baizik elkarrengandik bereiztea eragingo dute korronte (0,001642) sortu berriek . Zientzialarien (0,001261) esanean (0,000696) , hainbatean gertatu da hori Lurraren (0,002915) historian (0,000463) ; 200-250 milioi urtean (0,000505) behin , hain zuzen . Hortaz , Pangea baino lehenago beste elkartze eta bereizte batzuk gertatu ziren , nonbait , eta hemendik aurrera ere gertatuko omen dira . (OR) : Plaka tektoniko deritzen egituren (0,00495) zati (0,016393) dira gaur egun ezagutzen ditugun kontinenteak . Plaka (0,000836) tektoniko horiek mantuaren (0,00051) gainean (0,002203) flotatzen ari dira . Bolatxoak : New York Unibertsitateko taldeak (0,001692) egin zuen esperimentuaren (0,001263) eskema (0,000583) . Bolatxoak tangaren (0,000802) alde (0,006623) batean pilatzean , konbenkzio-korronteak bira ematen du .

IV Eranskina

Galdera adibideak

Gaien Aurkibidea

IV.a Sorturiko galderen adibide bat

Jarraian *Kontinente* testuaren inguruan sistemak sorturiko 10 galdera pisutsuenak erakusten dira. Hauek pisuaren arabera ordenatuta daude eta sistemaren irteera formatuan agertzen dira.

```
<?xml version="1.0" encoding="utf-8" ?>
<galderak>
<galdera pisua="0.037266699825946534" kasua="ABS">
<sortutako_galdera>
      [Zer] bete zuten Konbekzio-korrontea
         eratutakoan , hiru milimetroko bolatxoz ,
         bolatxoen eragina ikusteko $.?
</sortutako_galdera>
<jatorrizko_esaldia>
      Konbekzio-korrontea eratutakoan , hondoaren
         azalera erdia hiru milimetroko bolatxoz bete
         zuten , bolatxoen eragina ikusteko $.
</jatorrizko_esaldia>
<galderaren_emaitza>
      hondoaren azalera erdia
</galderaren_emaitza>
</galdera>

<galdera pisua="0.037266699825946534" kasua="INE">
<sortutako_galdera>
      [Non, Norengan, Noiz] bete zuten hondoaren
         azalera erdia hiru milimetroko bolatxoz ,
         bolatxoen eragina ikusteko $.?
</sortutako_galdera>
<jatorrizko_esaldia>
      Konbekzio-korrontea eratutakoan , hondoaren
         azalera erdia hiru milimetroko bolatxoz bete
         zuten , bolatxoen eragina ikusteko $.
</jatorrizko_esaldia>
<galderaren_emaitza>
      Konbekzio-korrontea eratutakoan ,
</galderaren_emaitza>
</galdera>

<galdera pisua="0.015176186235847856" kasua="ABS">
```

```
<sortutako_galdera>
      [Zer] berotu zuten azpian xafla bero bat
         jarrita?
</sortutako_galdera>
<jatorrizko_esaldia>
      azpian xafla bero bat jarrita berotu zuten
         likidoa $.
</jatorrizko_esaldia>
<galderaren_emaitza>
      likidoa $.
</galderaren_emaitza>

</galdera>

<galdera pisua="0.015176186235847856" kasua="INE">
<sortutako_galdera>
      [Non, Noiz] berotu zuten xafla bero bat jarrita
         likidoa $.?
</sortutako_galdera>
<jatorrizko_esaldia>
      azpian xafla bero bat jarrita berotu zuten
         likidoa $.
</jatorrizko_esaldia>
<galderaren_emaitza>
      azpian
</galderaren_emaitza>

</galdera>

<galdera pisua="0.007898068082477474" kasua="ABS">
<sortutako_galdera>
      [Zer] oztopatu zioten beroari $.?
</sortutako_galdera>
<jatorrizko_esaldia>
      beroari bidea oztopatu zioten $.
</jatorrizko_esaldia>
<galderaren_emaitza>
      bidea
</galderaren_emaitza>
</galdera>

```

```
<galdera pisua="0.00653230509628338" kasua="ABS">
<sortutako_galdera>
      [Zer] zuelako gaineko likido hotzagoak
         beherantz , handiagoa $.?
</sortutako_galdera>
<jatorrizko_esaldia>
      gaineko likido hotzagoak beherantz ,
         dentsitatea handiagoa zuelako $.
</jatorrizko_esaldia>
<galderaren_emaitza>
      dentsitatea
</galderaren_emaitza>
</galdera>

<galdera pisua="0.006032470173084171" kasua="ABS">
<sortutako_galdera>
      [Zer] gertatu da Zientzialarien esanean ,
         hainbatean Lurraren historian $; 200-250
         milioi urtean behin , hain zuzen $.?
</sortutako_galdera>
<jatorrizko_esaldia>
      Zientzialarien esanean , hainbatean gertatu da
         hori Lurraren historian $; 200-250 milioi
         urtean behin , hain zuzen $.
</jatorrizko_esaldia>
<galderaren_emaitza>
      hori
</galderaren_emaitza>
</galdera>

<galdera pisua="0.00603247017308417" kasua="INE">
<sortutako_galdera>
      [Non, Norengan, Noiz] gertatu da hainbatean
         hori Lurraren historian $; 200-250 milioi
         urtean behin , hain zuzen $.?
</sortutako_galdera>
<jatorrizko_esaldia>
      Zientzialarien esanean , hainbatean gertatu da
         hori Lurraren historian $; 200-250 milioi
         urtean behin , hain zuzen $.
</jatorrizko_esaldia>
```

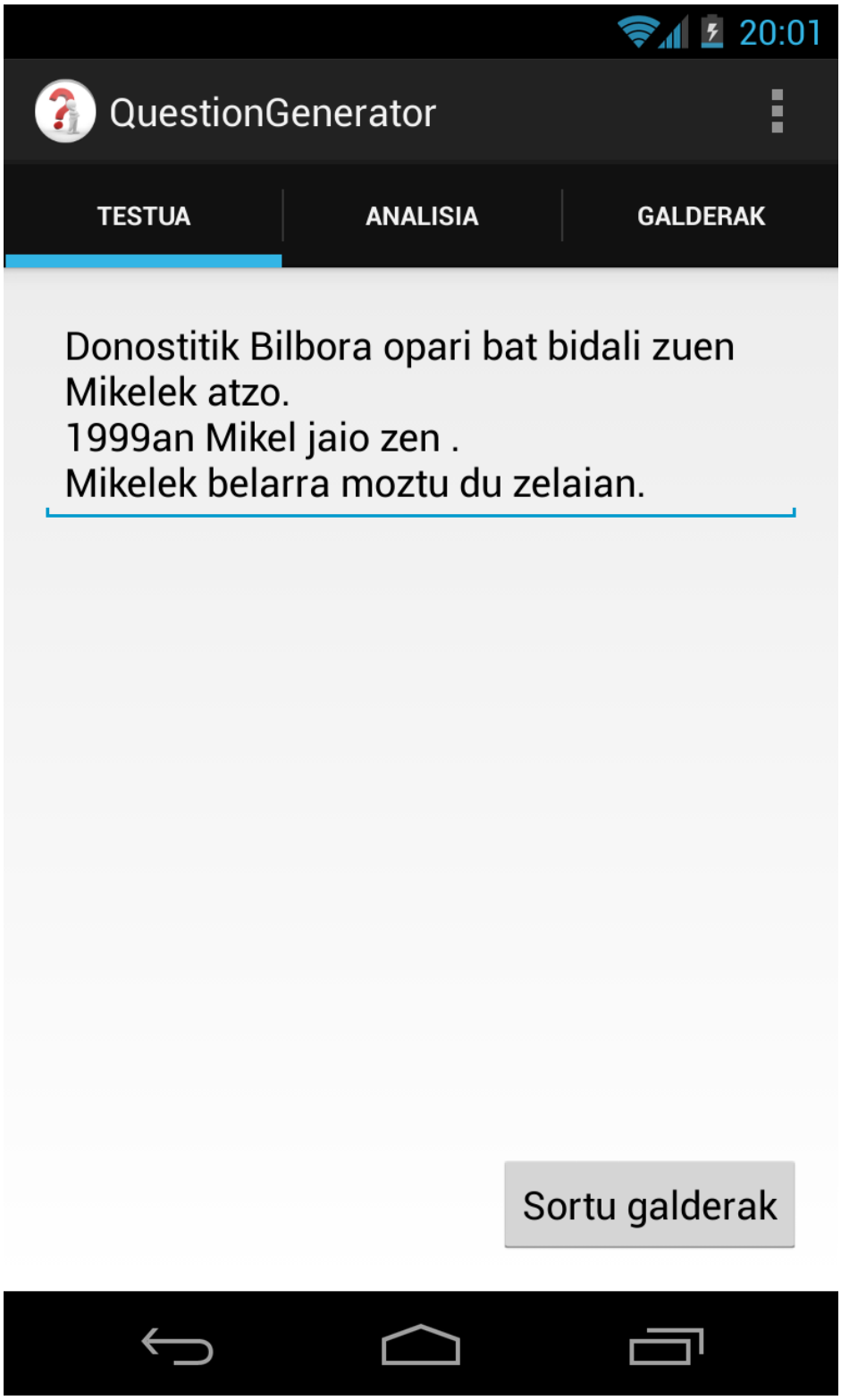
```
<galderaren_emaitza>
      Zientzialarien esanean ,
</galderaren_emaitza>
</galdera>

<galdera pisua="0.0037927868376165874" kasua="ABS">
<sortutako_galdera>
      [Zer] egin zuen esperimentuaren eskema $.?
</sortutako_galdera>
<jatorrizko_esaldia>
      New York Unibertsitateko taldeak egin zuen
         esperimentuaren eskema $.
</jatorrizko_esaldia>
<galderaren_emaitza>
      New York Unibertsitateko taldeak
</galderaren_emaitza>
</galdera>

<galdera pisua="0.0036747858837686094" kasua="ABS">
<sortutako_galdera>
      [Zer] egiten zuen gora dentsitatea galdutakoan
         ,?
</sortutako_galdera>
<jatorrizko_esaldia>
      azpialdean berotutako likidoak gora egiten zuen
         dentsitatea galdutakoan ,
</jatorrizko_esaldia>
<galderaren_emaitza>
      azpialdean berotutako likidoak
</galderaren_emaitza>
</galdera>

</galderak>
```

V Eranskina

Mugikorretarako integrazioa

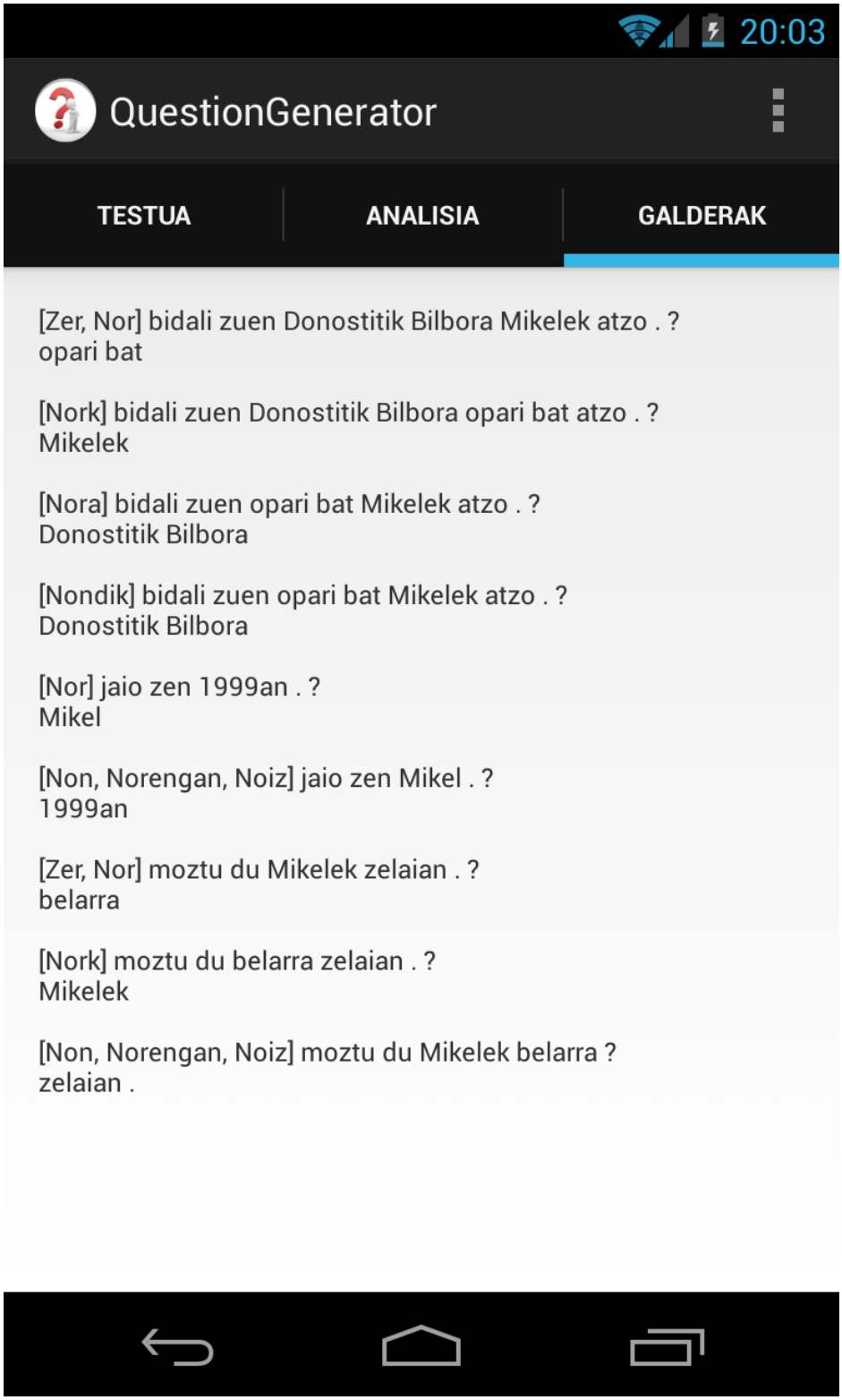
Gaien Aurkibidea

V.a Intefazeen irudi batzuk



Irudia V.1: Android integrazioa: testu sarrera interfazea

Irudia V.2: Android integrazioa: analisia ikusteko interfazea



Irudia V.3: Android integrazioa: galderak ikusteko interfazea

VI Eranskina

Erabilpen gida

Gaien Aurkibidea

Jarraian garatu diren hiru prototipoen erabilpen gidak aurkezten dira. Hiru aplikazioak Ixa ikerketa taldeko zerbitzarietan aurkitzen dira. Konkretuki honako helbidean: /sc01a4/users/jmadrazo003/proiektua/aplikazioak. Aurrerantzean erreferentziatzen diren direktorio eta aplikazio guztiak bertan aurkitzen dira. Aplikazioen kodeari dagokionez honako helbidean aurki daiteke: /sc01a4/users/jmadrazo003/proiektua/kodea.

Aplikazio hauen erabilpenari buruzko beste dudaren bat izanez gero bidali mezu bat jmadrazo003@ikasle.ehu.es helbidera.

VI.a QG-Malti galdera sortzaile automatikoa

QG-Malti galdera sortzailea exekutatzeko scriptak *QG-Malti.sh* izena du. Berau exekutatzeko sisx05 makinan egon beharra dago. Honela exekuta daiteke programa:

```
sh QG-malti.sh <fitxategia> <domeinu izena> <aukerak>
```

Lehen argumentua galderak sortzeko fitxategiaren helbidea da. Fitxategi honetan testuak esalditan banatuta egon beharko du. Hau da, lerroko esaldi bat. Hau ez da beharrezkoa izango etorkizun batean Maltixan esaldi banatzailea integratzen denean.

Bigarren argumentua domeinu izena da. Domeinu hau izango da galderen esangura neurtzeko erabiliko dena. Domeinuen informazioa *Jakintza* izeneko direktorioan gordetzen da eta termino erauzleen irteeratik sortu daiteke. Dokumentu hau idazten den momentuan bi domeinu daude erabilgarri, *teknologia* eta *historia*.

Bi argumentuen ondoren nahi adina aukera argumentu jar daitezke. Aukera posible guztiak gidoi ikur batez hasten dira. Hauek jarraian zerrendatzen dira:

- **-formatuaxmlpisugabe**
  Galderak XML formatuan inprimatuko dira. Galdera bakoitzaren pisua erakutsiko da, baina ez dira terminoen pisuak markatuko. Formatu hau Seneko plataformarekin bateragarria da eta zenbait ebaluazio bertan burutzeko erabilia izan da.

- **-formatuaxmlpisuekin**
  Aurreko formatuaren berdina da, baina galderen pisuak erakusteaz gain galderetako terminoak markatzen dira eta hauen pisua erakusten da.

- **-formatuadebug**
  Garapen fasean erabili da formatu hau, aplikazioa arazteko balio duelarik. XML formatuetan agertzen dena baino informazio gehiago ematen du. Irteera esaldika ordenatzen da eta esaldi bakoitzean analisi sintaktikoa erakusten da lehenik. Ondoren, bertatik sorturiko galdera guztiak erakusten dira. Galdera bakoitzean nodo hautagaia zein izan den adierazten da eta galdetzaile desanbiguaziorako aplikatu den teknika markatzen da. Ez ditu ordenazio aukerak errespetatzen, esaldikako ordenazio propio bat baitu.

- **-formatuaebaluazioa1**
  Galdera sortzaile automatikoaren ebaluazioan erabili zen formatu hau. Informazioa CSV formatuan adierazten da. Honek *Calc* eta *Excel* moduko aplikazioetan erabiltzeko erraztasunak ematen ditu. Galderen gramatikaltasuna neurtzeko beharrezko informazioa erakusten da bertan. Hala nola, galdera eta erantzuna.

- **-formatuaebaluazioa2**
  Galdera sortzaile automatikoaren ebaluazioan erabili zen formatu hau. Informazioa CSV formatuan adierazten da. Honek *Calc* eta *Excel* moduko aplikazioetan erabiltzeko erraztasunak ematen ditu. Galdetzaileen egokitasuna neurtzeko beharrezko informazioa erakusten da bertan. Hala nola, aukeraturiko galdetzailea, jatorrizko esaldia eta galderaren erantzuna.

- **-ordenapisua**
  Galderak zein ordenatan erakutsiko diren zehazten du arazteko formatuan izan ezik. Galderak beren pisuaren arabera ordenaturik inprimatuko dira.

- **-ordenakasuak**
  Galderak zein ordenatan erakutsiko diren zehazten du arazteko formatuan izan ezik. Galderak beren kasuaren arabera ordenatuko dira. Kasuak ordena alfabetikoan ordenatzen dira.

Sistemaren emaitza <*fitxategia*>*.galderak* izeneko fitxategian gordetzen da.

VI.b Termino erauzleak

Proiektuan inplementatu diren bi termino erauzle prototipoak modu berdinean funtzionatzen dute erabilpenari dagokionez. Hauen izenak *te1_termino_erauzlea.sh* eta *te2_termino_erauzlea.sh* dira, adieretan oinarrituriko termino erauzlea eta kontsentsuan oinarriturikoa, hurrenez hurren. Honela exekuta daitezke:

```
sh tx_termino_erauzlea.sh <corpusen direktorioa> <irteera direktorioa>
```

Lehen argumentua terminoak erauzteko corpusak gordetzen diren direktorioaren helbidea da. Direktorio honek antolaketa zehatz bat izan behar du. Bere barnean dauden direktorio bakoitzak corpus bat errepresentatzen du. Corpus honen izena direktorioaren berdina izango da eta bere barneko dokumentu guztiak Kyoto Annotation Framework (KAF) formatuan egon beharko dira.[1] Honen adibide bat *proba_corpusak* direktorioan ikus daiteke.

Bigarren argumentuan erauziriko terminoak zein sistemak sortzen dituen beste tarteko fitxategiak non gordeko diren adierazten da. Direktorio hau existitzen ez bada sortu egingo da. Bertan sistemen arkitekturan azaltzen diren fitxategi guztiak gordeko dira. Hala nola, maiztasun fitxategiak, relebantzia fitxategiak, kontsentsu fitxategiak eta azkenik termino fitxategiak.

Bi aplikazio hauek exekutatzeko siuc03 makinan egon beharra dago.

VI.c Funtzionamendua ikusteko prestaturiko probak

Aipaturiko aplikazioak modu errazean probatu ahal izateko dokumentu batzuk prestatuta utzi dira. Batetik galdera sorkuntzan probak burutu ahal izateko teknologia domeinuko 5 dokumentu utzi dira esalditan banatuta. Hauek *proba_testuak* izeneko direktorioan aurki daitezke. Bestalde, termino erauzleekin probak burutzeko 20 dokumentuko 3 corpus txiki prestatu dira. Hauek *proba_corpusak* izeneko direktorioan aurki daitezke.

Jarraian galdera sortzailea probatzeko burutu beharreko prozesua erakusten da.

[1] Fitxategi bat KAF formatura bihurtzeko sisx makinetako txt2kaf_eu.pl aplikazioa erabil daiteke.

```
ssh sisx05.si.ehu.es
cd /sc01a4/users/jmadrazo003/proiektua/aplikazioak
sh QG-Malti.sh proba_testuak/Artikoa.txt teknologia
   -formatuaxmlpisuekin
sh QG-Malti.sh proba_testuak/Kontinente.txt
   teknologia -formatuadebug
sh QG-Malti.sh proba_testuak/Planeta.txt teknologia
   -ordenakasuak
```

Bestetik termino erauzleak probatzeko honako prozesua jarrai daiteke:

```
ssh siuc03.si.ehu.es
cd /sc01a4/users/jmadrazo003/proiektua/aplikazioak
sh te1_termino_erauzlea.sh proba_corpusak/
   te1_erauzitakoa/
sh te2_termino_erauzlea.sh proba_corpusak/
   te2_erauzitakoa/
```

Bibliografia

[1] Euskarazko wikipedia. URL: http://eu.wikipedia.org/.

[2] Ixa ikerketa taldea. URL: http://ixa.si.ehu.es.

[3] Lur entziklopedia tematikoa. URL: http://www.euskara.euskadi.net/r59-luredir/eu/contenidos/informacion/directorio_enciclopedia/eu_euhistor/artikulu.html.

[4] Seneko. URL: http://ixa2.si.ehu.es/seneko/.

[5] Zt corpusaren webgunea. URL: http://www.ztcorpusa.net/.

[6] E. Agirre (1), X. Artola (1), A. Diaz de Ilarraza (1), G. Rigau(1), A. Soroa (1), and W. Bosma (2). Kaf: Kyoto annotation framework. *IXA group, University of the Basque Country (1), Computational Lexicology and Terminology Lab, VU Amsterdam (2).*

[7] Delphine Bernhard adn Louis de Viron, Veronique Moriceau, and Xavier Tannier. Question generation for french collating parsers and paraphrasing questions. 2012.

[8] I. Aduriz, M. Aranzabe, J.M. Arriola, A. Diaz de Ilarraza, K. Gojenola, M. Oronoz, and L. Uria. A cascaded syntactic analyser for basque. computational linguistics and intelligent text processing. *pp 124-135.*, 2004.

[9] Itziar Aldabe, Itziar Gonzalez-Dios, Inigo Lopez, Ion Madrazo, and Montse Maritxalar. Two approaches to generate questions in basque (forthcoming). *SEPLN*, 2013.

[10] Itziar Aldabe, Montse Maritxalar, and Ander Soraluze. Arikiturri. *Ixa ikerketa taldea*, 2011.

[11] I. Aldezabal. Aditz-azpikategorizazioaren azterketa. 100 aditzen azterketa zehatza, levin (1993) oinarri harturik eta metodo automatikoak baliatuz. *UPV/EHU*.

[12] Aitzol Astigarraga, Koldo Gojenola, Kepa Sarasola, and Aitor Soroa. *TAPE Testu-analisirako PERL erremintak*. Udako Euskal Unibertsitatea (UEU), Bilbo, Spain, 2009.

[13] Sergio Curto, Ana Cristina Mendes, and Luisa Coheur. Question generation based on lexico-syntactic patterns learned from the web. *Spoken Language Systems Laboratoy and University of Lisbon*, 2012.

[14] A. Diaz de Ilarraza, A. Mayor, and K. Sarasola. Semiautomatic labelling of semantic features. *UPV/EHU, 19th International Conference on Computational Linguistics*, 2002.

[15] Michael Heilman and Noah A. Smith. Good question! statistical ranking for question generation. *Language Technologies Institute from Carnegie Mellon University*, 2010.

[16] Lopez-Gazpio I. and Maritxalar M. Web application for reading practice. *IADAT: International Conference on Education (forthcoming)*, 2013.

[17] I.Alegria(1), A.Gurrutxaga(2), P.Lizaso(2), X.Saralegi(2), S.Ugartetxea(2), and R.Urizar(1). A xml-based term extraction tool for basque. *(1) Ixa taldea and (2) Elhuyar Fundazioa*, 2004.

[18] Elisabete Pociello Irigoyen. Euskararen ezagutza-base lexikala: Euskal wordnet. *Euskal Filologian Doktore titulua eskuratzeko aurkezturiko Tesia*, 2007.

[19] Bengoetxea K. and Gojenola K. Application of diferent techniques to dependency parsing of basque first workshop on statistical parsing of morphologically rich languages. *SPMRL and NAACL Workshop Los Angeles*, 2010.

[20] Ming Liu, Rafael A. Calvo, and Vasile Rus. G-asks: An intelligent automatic question generation system for academic writing support. *University of Memphis and University of Sydney*, 2012.

[21] Iñigo Lopez Gazpio. Seneko: galderak automatikoki sortuz testuak lantzeko aukera ematen duen aplikazioa. *Ixa ikerketa taldea*, 2013.

[22] Andreq M. Olney, Arthur C. Graesser, and Natalie K. Person. Question generation from concept maps. *University of Memphis and Rhodes College*, 2012.

[23] Andrew M. Olney, Arthur C. Graesser, and Natalie K. Person. Tutorial dialog in natural language. *University of Memphis and Rhodes College*, 2010.

[24] Bogdan Scaleanu Paul Buitelaar. Ranking and selecting synsets by domain relevance. *DFKI GmbH*, 2001.

[25] Vasile Rus and Arthur C Graesser. The question generation shared task and evaluation challenge. *The University of Memphis: National Science Foundation*, 2009.

[26] Paola Velardi, Michele Missikoff, and Roberto Basili. Identification of relevant terms to support the construction of domain ontologies. *Universita di Roma*, 2001.

Two Approaches to Generate Questions in Basque*

Dos aproximaciones para generar preguntas en euskera

Itziar Aldabe y **Itziar Gonzalez-Dios** y **Iñigo Lopez-Gazpio**,
Ion Madrazo y **Montse Maritxalar**
IXA NLP Group, University of the Basque Country UPV/EHU
Manuel Lardizabal Pasealekua 1 20018 Donostia
{itziar.aldabe, itziar.gonzalezd, ilopez077, jmadrazo003, montse.maritxalar}@ehu.es

Resumen: En este artículo se presenta un generador de preguntas basado en chunks y otro generador basado en dependencias sintácticas. Ambos generan preguntas en euskera a nivel de frase y utilizan el rasgo de animado/inanimado de los nombres, las entidades nombradas y los roles semánticos de los verbos, así como la morfología de los sintagmas nominales. Se describen dos experimentos de generación de preguntas basadas en textos didácticos, en los que una lingüista analiza la gramaticalidad y lo apropiado de las preguntas generadas a partir de frases simples, así como sus correspondientes pronombres interrogativos.
Palabras clave: generación de preguntas, recursos didácticos, rasgos semánticos, morfosintaxis

Abstract: This article presents a chunker-based question generator (QG) and a QG system based on syntactic dependencies. Both systems generate questions in Basque at the sentence level and make use of the animate/inanimate feature of the nouns, named entities, semantic roles of the verbs, and the morphology of the noun phrases. Two experiments to generate questions were carried out based on educational texts. Then, a linguist analysed the grammaticality and appropriateness of the questions generated from single sentences, as well as their interrogative pronoun.
Keywords: question generation, educational resources, semantic features, morphosyntax

1 Motivation

A new community of interdisciplinary researchers[1] have found a common interest in generating questions.[2] The first workshop on the question generation shared task and evaluation challenge (QGSTEC-2008) began the discussion on the fundamental aspects of question generation (QG) and set the stage for future developments in this emerging area. QG is defined (Rus and Graesser, 2009) as the task of automatically generating questions from some form of input, for which the input could vary from raw text to in-depth semantic representation.

Most of the current QG systems are mainly focused on the generation of questions based on single sentences (Rus and Graesser, 2009; Boyer and Piwek, 2010; Graesser et al., 2011). The generation task contains three steps (Rus and Graesser, 2009): content selection, question type selection, and question construction. The content identification and question type selection (i.e. interrogative pronoun) are usually carried out based on various linguistic information. This information is obtained by means of several natural language processing (NLP) tools: syntactic analysers, named entity recognisers, correference resolution systems and semantic role labellers. In contrast, the question construction is usually based on some transformation rules and patterns.

There is few research work on the generation of questions from paragraphs. An approach is presented in Mannem, Prasad, and Joshi (2010) where the generation is based on semantic roles of predicates. Agarwal, Shah, and Mannem (2011) also present

* Acknowledgments: we thank the linguist Ainara Estarrona for her help in the definition of the wh-word lists related to the semantic roles. This research was partially funded by the IRAKURRI project (SPE12UN091), the Ber2Tek project (IE12-333), and the SKaTeR project (TIN2012-38584-C06-02)

[1] Researchers from various disciplines such as cognitive science, computational linguistics, computer science, discourse processing, educational technologies and language generation.

[2] http://www.questiongeneration.org/

a system which generates questions based on more than one sentence. For that task, they use discourse connectives. Other researchers address the task of generating the questions from a more pedagogical or psychological point of view. For instance, Mostow and Chen (2009) present a system based on a situation model which is based on characters' mental states. More recently, Olney, Graesser, and Person (2012) generate questions from concept maps based on psychological theories.

Aldabe, Maritxalar, and Soraluze (2011) have probed the viability of the QG task for Basque language. They use a numerical entity recogniser and classifier to detect numerical entities and generate questions about them. Previous to the creation of the questions, the system automatically detects the clauses to be used for the generation. However, the present work deals with the automatic generation of Basque questions using single sentences as the source text for the generation.

This article presents two approaches to generate questions, a chunker-based generation and a generation based on syntactic dependencies. Both approaches created direct questions regarding the noun phrases of the sentences of the corpora. In general, we foresee a better performance when using syntactic dependencies. However, we expect that a chunker-based generation can also be suitable if we limit the source input to single sentences. As the final aim of the system is to be used in the education domain, authors evaluate both approaches whith texts prepared to work on science and technology at secondary school and texts prepared for a language learning scenario. The evaluation focused on how well each approach transforms a sentence into its corresponding interrogative form.

The paper is structured as follows. Section 2 presents the main features used for the generation process. Section 3 describes the implemented systems. Section 4 explains the results of the experiments. Finally, conclusions and future work are commented on section 5.

2 Question Generation

This work presents two question generation systems for Basque. The generation is based on the morphological information of the noun phrases. The systems also use semantic features during the generation process.

2.1 QG based on Noun Phrases

Basque is a Pre-Indo-European language and differs considerably in grammar from the languages spoken in surrounding regions. It is, indeed, an agglutinative head-final isolated language. The case system is ergative-absolutive. The inflections of determination, number and case appear only after the last element in the noun phrase. This last element can be the noun, but also typically an adjective or a determiner. Basque nouns belong to a single declension and its 18 case markers are invariant. Functions, normally fulfilled by prepositions, are realised by case suffixes inside wordforms.

In this work we intend to automatically generate questions about all the noun phrases appearing in each sentence. To that end, the detection of the case markers of the noun phrases is the starting point for the generation process. We report two QG systems in order to compare a chunker-based approach and a dependency-based approach.

We choose 5 case markers as the starting point for the experiments: absolutive (ABS), ergative (ERG), inessive (INE), allative (ALL) and ablative (ABL). As explained in section 4.1, all of them cover almost the 90% of all the noun phrases found in corpora when generating questions in our scenario. Absolutive and ergative cases accumulate the highest percentage of noun phrases in the corpus, as they are related to the subject and direct object syntactic functions. The inessive case is used in noun phrases with different adverbial functions (temporal, location, etc). And the ablative and allative cases give us the chance to work with the animate/inanimate features. The mentioned 5 cases need different wh-words (interrogative pronouns) depending on the features of the head in the noun phrase.

2.2 Semantic Features for QG

We have explored the animate/inanimate feature of the nouns, the use of named entities (person, location and organisation) and the semantic roles to deal with the generation of questions.

- Animate/inanimate feature: the QG generators use the work done by Díaz de Ilarraza, Mayor, and Sarasola (2002),

where semantic features of common nouns are extracted semi-automatically from a monolingual dictionary. Both systems consider the animate/inanimate feature of 15,000 nouns.

- Named entities: both QG systems include a named entity recogniser and classifier named *Eihera* (Alegria et al., 2003). They use this tool to identify person, place and organisation entities.
- Semantic roles: The QG generators take into account a corpus manually tagged at the predicate level with verb senses, argument structure and semantic roles (Aldezabal et al., 2010). This corpus is based on the work done in Aldezabal (2004), which includes an in-depth study of 100 verbs for Basque. Based on the occurrences of the 100 verbs, we have worked with the following roles from VerbNet (Kipper et al., 2006): actor, attribute, agent, beneficiary, cause, destination, direction, experiencer, extent, instrument, location, manner, patient, predicate, product, recipient, source, theme, temporal and topic. The mandatory roles of patterns with a probability higher than 75% are considered to be candidates.

3 QG-Malti and QG-Ixati

The article reports two question generation systems for Basque, QG-Malti and QG-Ixati. QG-Malti is a QG system based on *Maltixa* (Bengoetxea and Gojenola, 2010), a dependency parser for Basque. QG-Ixati is a QG system which uses *Ixati* (Aduriz et al., 2004), a chunker for Basque. Previous to the selection of the noun phrase (the content selection step), both systems perform a morphosyntactic analysis of the source texts.

As proposed in Rus and Graesser (2009), both QG systems can be described as a three-step process: content selection, question type selection and question construction. The main constraint on the present study is that the systems only select sentences which contain a single finite verb. Before the content selection process, QG-Malti also splits coordinate sentences into single sentences.

The goal of both approaches is to generate questions at sentence level. QG-Malti discards the sentences which have discourse elements whose function is to connect the sentence with other elements outside the sentence, but it rejects them only in case the discourse elements are not at the beginning of the sentence. QG-Ixati, however, can not discard this kind of sentences as the analyser Ixati does not detect this kind of discourse relations.

3.1 Target Selection

As mentioned, both systems generate questions related to all the noun phrases that occur in the sentences of the source text[3]. The generation process uses morphosyntactic features of the output of the corresponding analyser to select **the candidate target**. In the case of the QG-Ixati, the candidate target is the whole noun phrase (chunk). However, in the case of QG-Malti the candidate target is the word whose morphological analysis has the target case marker. And then, the dependency structure of the analysis is used to construct the corresponding noun phrase.

When there is more than one occurrence for the same case marker inside the same sentence, only one of those occurrences is used to generate a question. Based on the fact that in Basque the relevant information of a sentence is close to the verb, QG-Malti selects as the candidate target (word) the one which is closest to the verb. And, if there are two candidates at the same distance to the verb, it selects the one located on the left to the verb. The reason for this criterion is that in Basque the informationally relevant phrase of a sentence precedes immediately the verb.

QG-Ixati, however, establishes a preference criterion based on various semantic features of the candidate targets (noun phrases). It gives a higher priority to the animate/inanimate feature and named entity tag than to the semantic roles. The priority is obtained as follows:

1. If the head of the noun phrase is a named entity (person, place or organisation) or its animate/inanimate feature is known, the QG system establishes a weight of 2 for the noun phrase.
2. If the noun phrase fulfills one of the mandatory roles of a particular verb subcategorization pattern, the QG system establishes a weight of 1 for the given noun phrase.

[3]We use 5 declension cases in the experiments of the present work.

	Animate	Person	Inanimate	Place	Organisation	No semantic feature
ABS	*Nor*		*Zer*			*Nor/Zer*
ERG	*Nork*		*Zerk*			*Nork/Zerk*
INE	*Norengan*		*Non*			*Norengan/Non/Noiz*
ALL	*Norengandik*		*Nondik*			*Norengandik/Nondik*
ABL	*Norengana*		*Nora*			*Norengana/Nora*

Table 1: Question type based on named entities, animate/inanimate and case markers. Nor (Who-ABS); Zer (What-ABS); Nork (Who-ERG); Zerk (What-ERG); Norengan (To whom); Non (Where); Noiz (When); Norengandik (From whom); Nondik (From where); Norengana (To whom); Nora (To where)

The system chooses the noun phrase with the highest priority. In the cases that the system still assigns the same weight to different noun phrases, the selected candidate is the one which is closest to the verb. And in case of still being a tie, the system chooses the phrase located on the left to the verb.

3.2 Question Type Selection

QG-Malti and QG-Ixati follow the same criteria when selecting the question type to be generated. The selection of the question type is based on the linguistic information of the corresponding candidate target. For each case marker and linguistic feature (animate/inanimate, named entity, semantic role and morphology), an expert in the field established the most probable question type (wh-word) based on linguistic studies, as well as on her experience.

Table 1 shows the question types selected by the QG systems related to the named entity, animate/inanimate feature and case marker of the candidate target. For example, if the head of the noun phrase is identified as a person named entity and its corresponding case marker is the absolutive, the *NOR* (Who-ABS[4]) wh-word is selected.

The question type is also selected based on the semantic role of the candidate target. In total, 11 different roles have been linked to targets with the absolutive case, 7 to the ergative, 8 to the inessive, 5 to the ablative, and 5 to the allative. Depending on the semantic role of the candidate target, the QG system establishes its corresponding wh-word. For each verb, mainly only one question type is linked to each role. But, there are some exceptions, for example, the verb *compare* can have an animate or inanimate *patient* that correspond to the *NOR* (Who-ABS) and *ZER* (What-ABS) question types respectively.

3.3 Question Construction

In this phase, each QG system applies its own strategy based on the information given by the corresponding analyser. QG-Malti constructs the questions using the dependency relation structure analysed in the source sentence. QG-Ixati uses the information of the chunks detected during the morphosyntactic analysis of the source sentence.

The question building is based on simple transformation rules defined in the system. The first element of the constructed question is the wh-word. Following the wh-word, the main verb is established. Then, the rest of the elements (dependency structures or chunks) that are to the left of the verb in the source sentence are added to the question. Finally, the elements that appeared on the right of the source sentence's verb are appended to the generated question.

During the development of the systems we realised that some discourse connectives (e.g. the connective *gainera* [5]) caused some noise to the generated questions. In most of the cases where the connective was at the beginning of the source sentence, such a noise could be avoided if the connective was deleted when constructing the question. That is why we decided to delete from the source sentences all the discourse connectives which appear at the beginning of the sentence.

4 Evaluation

For the experiments, we chose texts about science and technology for secondary school

[4]The ABS mark refers to the fact that the wh-word takes the absolutive case marker.

[5]Basque word for *in addition*

learners and a specialised corpus in language learning because one of the final aims is to use QG systems into the education domain. In this work, as a first step, the evaluation focused on how well each QG system transforms a sentence into its corresponding interrogative form.

We focused on the evaluation of the syntactic correctness and fluency of the generated questions. To do so, a human judge followed the same classification as the one proposed in Boyer and Piwek (2010). We also studied the quality of the question types determining whether the generated wh-words asked about the source sentence. Finally, the expert also established whether the question was appropriate in relation to the source sentence.

4.1 Datasets

The science and technology (ST) dataset is composed of 5 texts about science and technology. One expert who works on the generation of learning materials defined the 5 texts as adequate for secondary school learners (Aldabe and Maritxalar, 2010). The main topics of these texts were: Continent; the Earth; Bats; the Arctic; and Computers respectively. All the texts have a similar length. In total, the dataset contains 176 sentences, being the average length of a sentence 13 words.

The language learning (LL) dataset focuses on a specialised corpus for Basque language learning, which is a collection of learning-oriented Basque written texts. The corpus is classified into different language levels[6] in accordance with the Common European Framework of Reference for Languages (Little, 2011). In the present work, the intermediate level of the corpus is the basis to generate the questions. The corpus is composed of near 80,000 sentences (over one million words), and the average length of a sentence is 13 words.

The ST dataset contains 646 noun phrases and the LL dataset has 200,000 noun phrases. Looking at the 5 case markers that are the starting point of the systems to generate the questions, almost 90% of the noun phrases are covered with the mentioned target case markers in both datasets. In the ST dataset, 55% of the noun phrases have the absolutive case marker, the 12% of the phrases have the ergative case marker, the 16% of phrases are inessive, the 3% of noun phrases have the allative case and the 2% of them the ablative case. In the LL dataset, the 60% have the absolutive case marker, the 11% of the phrases have the ergative case marker, and the 16% of phrases are inessive. Regarding the allative and ablative cases the percentage is near the 3%. The rest of case markers are under the 4% in both datasets.

4.2 Experiments

For each dataset, experiments with both QG systems were performed. The questions generated by QG-Malti and QG-Ixati were manually evaluated at different levels by one linguist.

As regards the question-types, a linguist judged whether the generated wh-words asked about the source sentence (yes/no). For that, the source sentence (input for the QG system) and the candidate target (answer to the generated question) were provided.

When checking the grammaticality, the linguist evaluated the syntactic correctness and fluency of the generated questions. For that, only the generated questions were provided. The questions were classified and differentiated among: i) correct questions; ii) questions which need minor changes (punctuation, capitalization, spelling or dialectical variants); iii) questions with major changes that are unnatural for native speakers even they are grammatically correct; and iv) incorrect questions due to the grammar, including oral speech style.

Finally, the judge established if the generated questions were appropriate (yes/no). For that, in addition to each question, the corresponding answer was also shown. When evaluating the appropriateness of the questions only correct questions and questions which nedded minor changes were considered.

4.2.1 ST dataset experiment

The experiment with the ST dataset reflects an educational scenario where the creation of updated material using texts from the web is crucial for the motivation of learners and teachers. Both systems generated questions for all the candidate targets of the 5 texts. Based on the 5 case markers, QG-Malti and

[6] Although the language level of a text can be a controversial aspect because it is difficult to define, in our source corpus, expert teachers classified the texts into specific levels.

	ST-common			ST-divergent		
	Wh-word	**Grammar**	**Appropr.**	**Wh-word**	**Grammar**	**Appropr.**
QG-Malti	76%	54%	46%	72.6%	59.7%	41.9%
QG-Ixati	88%	66%	64%	87.1%	48.4%	48.4%

Table 2: Results for the ST common and divergent inputs

QG-Ixati generated 112 and 81 questions respectively.

	Wh-word	**Grammar**	**Appropr.**
QG-Malti	75.0%	57.1%	44.6%
QG-Ixati	87.6%	59.3%	58.0%

Table 3: Percentage of correct questions of the QG systems for the ST dataset

Table 3 presents the evaluation results as regards wh-words, grammaticality and appropriateness in the ST dataset. The grammar column groups questions marked as correct and questions which need minor changes. In general, QG-Ixati obtains better results than QG-Malti, but, QG-Malti generates more questions. Thus, QG-Malti generates 64 grammatically correct questions while QG-Ixati generates 48.

The generation processes of QG-Malti and QG-Ixati differ mainly due to the analysers and the target selection criteria. However, both systems have in common some instances. We refer to common instances to those which have the same candidate target with the same case marker. Out of the 112 and 81 generated questions both systems have in common 50 instances. Thus, apart from the these common instances, QG-Malti selects 62 sentences to generate the questions, while QG-Ixati chooses other 31 different ones. Table 2 presents the manual evaluation results based on this distinction. As regards the common instances (ST-common column), QG-Ixati obtains better results in terms of wh-words (88%), grammaticality (66%) and appropriateness (64%). The comparison of the divergent samples (ST-divergent column) with the common instances of each system shows different results. On the one hand, it is remarkable the improvement of the grammaticality of QG-Malti (59.7%) compared to its common instances (54%). On the other hand, QG-Ixati obtains worse results in terms of grammaticality (48.4%) and appropriateness (48.4%), compared to the common instances (66% and 64% respectively).

Thus, even the overall results are better for QG-Ixati, the number of grammatically correct questions of the divergent dataset is higher in the case of QG-Malti. These results must be analysed deeply as we foresee that the target case markers and the used analysers can have an influence on the results.

4.2.2 LL dataset experiment

The aim of the LL dataset experiment is to analyse the influence of the case marker of the noun phrase chosen as the answer to the generated question. This is why the sample contains 20 questions per case marker for each QG system selected at random[7]. In this experiment, a total of 100 generated questions for each system are evaluated.

Table 4 shows the evaluation results per case marker. In general, both systems obtain grammatically better questions when the generation is based on noun phrases with absolutive or inessive case markers. QG-Ixati obtains better overall results compared to QG-Malti. It is noticeable the difference on the grammaticality of the absolutive (QG-Ixati, 85% and QG-Malti, 60%) and ergative (QG-Ixati, 65% and QG-Malti, 45%) case markers. In contrast, QG-Malti performs better in terms of grammaticality and appropriateness of the allative and ablative case markers, and the wh-word of the ergative.

Although the source sentences are the same for both systems, the systems sometimes differ in the source candidate targets for the generation process. Out of the 100 questions, both systems have in common 47 questions. Table 5 presents the results of the 100 questions (LL-overall column) and the 47 common questions (LL-common column) in terms of wh-word, grammaticality and appropriateness.

The grammatically is better for QG-Malti when looking at the common instances (from 57% to 63.8%) and it is lower for QG-Ixati

[7]The source sentences were the same for both QG systems.

	QG-Malti			QG-Ixati		
	Wh-word	**Grammar**	**Appropr.**	**Wh-word**	**Grammar**	**Appropr.**
ABS	70%	60%	55%	85%	85%	60%
ERG	95%	45%	40%	80%	65%	60%
INE	60%	85%	60%	70%	85%	60%
ALL	70%	45%	35%	85%	35%	30%
ABL	50%	50%	45%	55%	40%	35%

Table 4: Percentages per case markers (20 questions per case marker)

	LL-overall			LL-common		
	Wh-word	**Grammar**	**Appropr.**	**Wh-word**	**Grammar**	**Appropr.**
QG-Malti	69%	57%	47%	70.2%	63.8%	57.4%
QG-Ixati	75%	62%	50%	68.1%	59.6%	48.9%

Table 5: Results for the LL overall and LL common inputs

(from 62% to 59.6%). Looking at the case markers of the 47 questions, just 3 out of the 47 questions correspond to the absolutive noun phrases and this is the main reason for getting worst results when using QG-Ixati.

4.3 Preliminary Error Analysis

The analysis of the results as well as the subsequent meetings with the expert allowed us to carry out a preliminary error analysis of the systems.

As regards the grammatical correctness of the questions, we have classified the erroneous questions in different categories: (i) questions which are grammatically correct but unnatural as regards the speakers; (ii) questions which contain orthographic errors; (iii) questions which are incorrectly generated in terms of morphology; (iv) questions which refer to oral speech; (v) problems with punctuation marks; and (vi) questions with an incorrect word order.

One of the reasons to generate ungrammatical questions is due to the type of the source input. In the analysis of the results we detected: i) some source input that correspond to subordinate clauses; ii) some source input that correspond to relative clauses and iii) some typos or spelling errors at word level. When analysing the results without taking into account the mentioned questions, the grammaticality and appropriateness measures of both QG systems improve 6 points for the ST dataset, and more than 10 points for the the LL dataset. In contrast, the number of correct wh-words hardly varies.

5 Conclusions and Future Work

Our QG systems created questions in order to ask about noun phrases at sentence level. With that end, a chunker-based QG system, QG-Ixati, and a QG system based on dependency structures, QG-Malti, have been implemented. Both systems deal with Basque language and make use of the animate/inanimate feature of the nouns, named entities (person, location and place), the semantic roles of the verbs, as well as the morphology of the noun phrases.

The results of the experiments show that QG-Malti generates a higher number of questions in a real scenario (ST dataset), however its general performance is slightly worse than QG-Ixati. The results for the LL dataset show a noticeable difference in grammaticality between both systems when generating questions about noun phrases with the absolutive and ergative case markers.

Future work will focus on the improvement of the systems. Once the roles are detected automatically, the semantic role approach would cover more verbs. Thus, we plan to focus on the analysis and integration of new Basque NLP tools or knowledge representations in order to generate questions that require deeper understanding. In addition, in the case of QG-Malti we want to improve the system using the information about syntactic dependencies, to discard ungrammatical source input for the generator, and to im-

prove the results of the identification of the question type.

Bibliography

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

Agarwal, Manish, Rakshit Shah, and Prashanth Mannem. 2011. Automatic question generation using discourse cues. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics.

Aldabe, Itziar and Montse Maritxalar. 2010. Automatic distractor generation for domain specific texts. In *Proceedings of the 7th International Conference on NLP, IceTAL 2010*. Springer, pages 27–38.

Aldabe, Itziar, Montse Maritxalar, and Ander Soraluze. 2011. Question generation based on numerical entities in Basque. In *Proceedings of AAAI Symposium on Question Generation*, pages 2–8.

Aldezabal, Izaskun. 2004. *Aditazpikategorizazioaren Azterketa Sintaxi Partzialetik Sintaxi Osorako bidean. 100 aditzen azterketa, Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz*. Ph.D. thesis, Euskal Filologia Saila. UPV/EHU.

Aldezabal, Izaskun, María Jesús Aranzabe, Arantza Díaz de Ilarraza, Ainara Estarrona, and Larraitz Uria. 2010. EusPropBank: Integrating semantic information in the Basque dependency treebank. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 60–73.

Alegria, Iñaki, Nerea Ezeiza, Izaskun Fernandez, and Ruben Urizar. 2003. Named Entity Recognition and Classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperación de Información, JOTRI, Madrid.*

Bengoetxea, Kepa and Koldo Gojenola. 2010. Application of different techniques to dependency parsing of Basque. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 31–39. Association for Computational Linguistics.

Boyer, Kristy Elizabeth and Paul Piwek, editors. 2010. *Proceedings of QG2010: The Third Workshop on Question Generation*. Pittsburgh: questiongeneration.org.

Díaz de Ilarraza, Arantza, Aingeru Mayor, and Kepa Sarasola. 2002. Semiautomatic labelling of semantic features. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Graesser, Arthur, James Lester, Jack Mostow, Rashmi Prasad, and Svetlana Stoyanchev. 2011. Question generation papers from the AAAI fall symposium. Technical report, FS-11-04.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *In Proceedings of 5th international conference on Language Resources and Evaluation*.

Little, David. 2011. The common european framework of reference for languages: A research agenda. *Language Teaching*, 44(03):381–393.

Mannem, Prashanth, Rashmi Prasad, and Aravind Joshi. 2010. Question generation from paragraphs at UPenn: QGSTEC system description. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 84–91.

Mostow, Jack and Wei Chen. 2009. Generating instruction automatically for the reading strategy of self-questioning. In *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, pages 465–472. IOS Press.

Olney, Andrew M, Arthur C Graesser, and Natalie K Person. 2012. Question generation from concept maps. *Dialogue & Discourse*, 3(2):75–99.

Rus, Vasile and Arthur C. Graesser, editors. 2009. *The Question Generation Shared Task and Evaluation Challenge*.