eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Informatika Ingeniaritzako Gradua

Konputazioa

Gradu Amaierako Proiektua

Rol semantikoen integrazioa TectoMT itzultzailean

Egilea

Oneka Jauregi Mikelarena

informatika fakultatea
facultad de informática

2016

Laburpena

Gaur egun ohikoa da hizkuntza batetik bestera pasatzen laguntzeko itzultzaile automatikoak erabiltzea. Hala ere, gauza jakina da pertsona batek egindako itzulpenaren kalitatetik urrun gelditzen direla itzulpen hauek. Ildo honi jarraituz, proiektu honek, hizkuntzalaritza eta informatika jorratzen ditu.

Jarraian aurkezten den proiektuan, rol semantikoak erabili dira, transferentzia bidezko TectoMT itzultzailean hobekuntza bat ekartzeko, ingelesetik euskararako itzulpenean. Hau da, hitz bakoitzak esaldiko esanahian jokatzen duen papera kontuan hartzeak, itzulpena hobetzen laguntzen duen ala ez ikusi nahi izan da.

Gaien aurkibidea

Irudien aurkibidea

Taulen aurkibidea

1. KAPITULUA

Proiektuaren deskribapena

Kapitulu honetan poiektuaren sarrera aurkezten da. Lana kokatzen den Hizkuntzaren prozesamenduaren arloa zer den azalduko da, eta honen barne dagoen itzulpen automatikoa ere bai.

1.1 Hizkuntzaren prozesamendua

Pertsonek gero eta gehiago erabiltzen dituzte makinak eta hauei egitea eskatzen zaien lan kopurua handituz doa. Esaterako, ohikoa da makinek hizkuntzarekin lotutako zereginetan laguntzea, batetik bestera pasatzeko, edo testu bat idatzi eta zuzentzeko. Konputagailu bat ez da gizakion lengoaia ulertzeko gai ordea, eta zaila da egiteko horiek pertsona batek bezain ongi burutzea.

Informatikariek Hizkuntzaren Prozesamendua (NLP, Natural Language Processing) deitzen duten ikerketa lerroak, arazo mota hauek hizkuntzaren tratamendu automatikoaren arloko teknikekin konpontzeko aukera ematen du. Hizkuntzalariek berriz Hizkuntzalaritza Konputazionala deitzen dute, nahiz eta zehazki berdina ez izan, bakoitzak bere arloa gehiago sakontzen baitu.

Dokumentu honetan agertuko diren kontzeptuak hobeto ulertzeko, ezagutza mota hauek barneratzea lagungarria da, nahiz eta gu semantika aldean gehiago zentratuko garen:

- Fonetikoa eta fonologikoa. Hitzak nola ahoskatu behar diren zehazten dute eta hizki bakoitzari zein fonema dagokion.

- Lexikala. Hizkuntzan erabil daitezkeen morfemak zehazten dira (lemak, aurrizkiak eta atzizkiak), eta bakoitzarentzat bere hizkuntza-ezaugarriak zehazten dira.
- Morfologikoa. Hitzen osaketa posibleak definitzen dira morfemak erabiliz.
- Sintaktikoa. Esaldien osaketa posibleak definitzen dira hitzak erabiliz.
- Semantikoa. Hitzen esanahia lortu eta osatzen duten esaldiaren esanahia lortzen da.
- Pragmatikoa. Esaldi baten interpretazio ezberdinak bereizten dira testuinguruaren arabera. Bi ataletan banatzen da:
  - Diskurtsoaren ezagutza. Izenorde, elipsi eta denbora-aspektuen interpretazioa. Elkarrizketako parte-hartzaile bakoitzak besteek dakitenari buruz edo nahi dutenari buruz suposatzen duena ere jakin behar da.
  - Munduaren gaineko ezagutza. Elkarrizketako gaiari buruz jakin beharrekoa da.

Hizkuntza automatikoki prozesatzeko tresnak errealitatea dira gaur egun. Badira zenbait aplikazio eskuragarri arlo honetan, hala nola, ortografia-zuzentzaileak, itzulpen automatikoa, itzulpen-laguntzak, hizketa testua bihurtzen duten sistemak, testua irakurtzen dutenak, hizkuntzak ikasteko sistemak eta abar luze bat.

Hala ere, guztietan, erronka nagusietako bat anbiguotasuna da, hau da, hitz batek esanahi bat baino gehiago izan ditzake. Testuinguruaren arabera, gizakiok esanahi egokia zein den ulertzeko gai gara. Konputagailuarentzat berriz, ez da hain lan erraza. Bestalde, hitzak ezin dira banan-banan tratatu, mendekotasunak baitaude esaldiko egituran. Esaldi batean hitz bat aldatuz, ez da esanahia bakarrik aldatzen, egitura osoa alda daiteke.

(1) a. Gizon bat ikusi nuen egunkariarekin.
    b. Gizon bat ikusi nuen teleskopioarekin.

(1-a) esaldian erraza da jakiten ikusitako gizonak egunkaria duela eskuan. (1-b) esaldian aldiz, gizonak teleskopio bat izan dezake, edo ikusleak teleskopioa erabili ahal izan du gizona ikusteko.

Lan honek itzulpen automatikoa du oinarritzat, hau da, makinaz soilik, hizkuntza batean idatzi bat emanda, beste hizkuntza batean lortzea. Hizkuntzaren prozesamenduan kokatzen den arlo hau azalduko da jarraian.

1.2 Itzulpen automatikoa

Proposamena 50. hamarkada arte ezagutzera eman ez zen arren, 1930. hamarkadan sortu zen lehen itzultzailea. Konputagailu bidezko lehen itzultzailea publikoki 1954an aurkeztu zen eta arrakasta handia izan zuen. Hala ere, 250 hitz bakarrik zituen eta esaldi zehatz batzuk besterik ez zituen itzultzen, errusieratik ingelesera.

Ikertzaileek, itzulpen automatikoaren arazoa hiru eta bost urte artean ebatziko zela sinesten zuten. Honek diru-laguntzak bultzatu, eta ikerketek aurrera jarraitu zuten. Sistemek hiztegi elebidunak eta erregela asko erabiltzen zituzten, eta gune batzuetan erabiltzen hasi ziren, baina bi esanahi dituzten hitzekin asmatzea adibidez ezinezkoa zen.

Estatu batuar eta sobietarrek ikerketak bertan behera utzi zituzten 60. hamarkadan, aurrerapena eskasa zela eta. Beste herrialdeetan aldiz hizkuntza ezberdinetako itzultzaileak garatzen hasi ziren, Kanada, Europa edo Japonian esaterako.

Geroztik, 80. hamarkadatik hasita, itzultzaileak gero eta gehiago garatzen joan dira, baita konputagailuak ere, hauen kostua merkatzearekin batera. Itzulpenak egiteko metodo ezberdinak erabili dira, eta pixkanaka hauen arteko konbinaketak ere bai.

Honen erakusle nabarmena dira 80 eta 90. hamarkadetan aurrera eramandako bi proiektu garrantzitsu: Eurotra [Hutchins et al., 1992] eta Verbmobil [Kay et al., 1992]. Lehenak Europar Batasuneko hizkuntza nagusien arteko itzulpena garatzea zuen helburu, transferentzia bidez. Verbmobil proiektuaren helburua berriz, Alemania hizkuntzen teknologian buru jartzearekin batera, itzultzaile mugikor bat sortzea zen, zuzeneko hizketa itzuliko zuena.

Konputagailu bidezko lehen itzultzaileek informazio teoria, mundu gerrako kriptografia eta lengoaia naturalaren oinarriak erabiltzen zituen.

Esan bezala, ondotik hiztegi elebidunak eta erregelak erabiltzen ziren. Hauek hitzen ordena zuzentzen laguntzen zuten, nahiz eta itzulpena asko baldintzatu. 80. hamarkadatik aurrerako teknologien garapenei esker, itzultzeko metodoak ere ugaritu eta hobetu ziren.

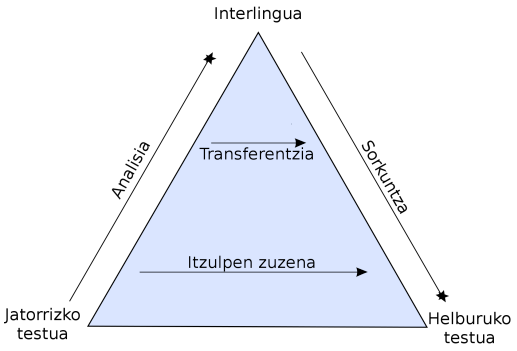
Garai hartan analisi morfologiko, sintaktiko eta semantikoa erabiltzen hasi ziren. Metodo estatistikoak garatu ziren 90. hamarkadan, baita adibideetan oinarritutako itzulpena ere. Bi metodo hauetan hobekuntza handiak izan dira urte hauetan. Hizketa-automatikoko itzultzaileekin ere lana egiten ari da. Hala ere askotan metodo batekin bestearen onurak ez ziren kontuan hartzen. Horregatik, sistema hibridoak ere garatu dira, metodo bat baino gehiago erabiltzen dituztenak. Ondoko puntuetan, metodo ezberdinak azaltzen dira:

1. Erregeletan oinarritutako itzulpenean jatorrizko eta helburuko hizkuntzen ezagutza linguistikoak erabiltzen dira. Hitz bakoitza itzultzeko hiztegi bat behar da, eta hizkuntza bata eta besteko egitura irudikatzeko erregela batzuk. Ondotik, bien arteko lotura egiten da.

   Hiru multzo bereizten dira metodo honetan:

   - Batetik, itzulpen zuzena, hiztegietan oinarritzen dena, oinarrizko erregela batzuekin itzultzen duena, pauso bakar batean.
   - Bestetik, transferentzia bidezko itzulpena. Hobeto azalduko da ondotik, lan honetan erabilitako sistemak metodo hau erabiltzen baitu. Itzulpena hiru fasetan egiten da: analisia, transferentzia eta sorkuntza. Bi adierazpen proposatzen dira, bat hizkuntza bakoitzeko. Hauetan esaldiaren informazio linguistiko zehatza gordetzen da.
   - Azkenik, Interlingua dago; honek hizkuntzekiko independentea den tarteko adierazpen bat sortzen du, itzulpena bi fasetan egitea ahalbidetuz: analisia eta sorkuntza.

   Jarraian 1.1 irudian erregeletan oinarritzen diren hiru itzulpen sistema hauen ezberdintasunak piramide moduan ageri dira. Irudi hau, *Vauquois* triangelua deitua, oso ezaguna da; itzulpen automatikoaren oinarriei buruzko artikulu gehienetan aurki daiteke.



1.1 Irudia: *Vauquois* hirukia

2. Adibideetan oinarritutako itzulpenean, corpus elebidunak (egituratutako testu bilduma handiak) erabiliz analogia egiten da. Itzulitako esaldiak gordetzen dira bertan, eta esaldi berri bat itzultzeko, corpusean egitura antzekotasunak dituzten esaldiak

aukeratzen dira. Antzeko esaldi zatiak itzuli eta elkartu egiten dira esaldi berria osatzeko.

3. Metodo estatistikoetan oinarritzen diren itzulpenak dira azken urteetan arrakasta handiena izan dutenak (SMT, S*tatistical Machine Translation*). Gizakiek itzulitako corpus elebidun handiak entrenatu eta ikasi egiten du sistemak, datu estatistikoak ateraz. Itzulpen eredua definitzen da: helburuko hizkuntzako kate bat, jatorrizko hizkuntzako kate bati dagokiona izateko probabilitatea. Helburua probabilitate hori maximizatzen duen kate parea aurkitzea da. Beraz, corpusean gero eta testu gehiago egon, orduan eta kalitate handiagoko itzulpena lortuko da.

   Erregeletan oinarritutako sistemek baino kostu txikiagoa izan ohi dute, corpusen beharra besterik ez baitago. Gainera, lortutako itzulpenak gehiago hurbiltzen dira lengoaia naturalera, gizakiek egindako itzulpenetatik hartzen baita testua. Hala ere, hizkuntza batzuetan corpus horiek lortzea zaila izaten da.

   Nahiz eta orain arte emaitza onenak metodo honekin lortu diren, mugak dituela ere ikusi da, eta aurrera pauso gutxi eman dira azken aldian. Akats zehatz batzuk zailak dira aurreikusi eta zuzentzen, eta itzulpena nahiko txarra izaten da hitzen ordenan ezberdintasun asko dituzten hizkuntzen artean. Azken 20 urteetan hutsetik hasi eta hobekuntza handiak lortu diren arren, gaur egun sabai batera heltzen ari dela uste da.

4. Sistema hibridoak, itzulpen automatikorako sistema bakarrean hurbilpen ezberdin bat baino gehiago erabiltzen dituztenak dira. Honen helburua, itzulpen metodo bakoitzaren alde onak batuz emaitza hobeak lortzea da, batak bestearen alderdi ahulak konpontzen lagunduz. Hainbat adibide daude sistema hauek ulertzeko:

   - Itzulpen bat baino gehiago paraleloan exekutatzea. Irteera testua, irteera guztien konbinaketa bat da. Orokorrean erregeletan eta estatistikan oinarritutako sistemak erabiltzen dira, baina beste multzo batzuk ere landu izan dira. Esate baterako, adibideetan, transferentzian, ezagutzan eta estatistikan oinarritutako itzultzaileak elkartu izan dituzte Estatu Batuetako *Carnegie Mellon* unibertsitatean (*Multi-Engine* [Jayaraman et al., 2005]).
   - Beste hurbilpen bat estatistika baliatuz erregelak sortzea da. Sarrera testua erregeletan oinarritutako ohiko itzultzaile baten gisara prozesatzen da erregela hauek baliatuz [Leja et al., 1998]. Domeinu espezifikoko testuekin nahiko emaitza onak ematen ditu.

- Azkenik, behin baino gehiagotan egiten den itzulpena ere izan daiteke sistema hibrido bat. Teknika hedatuena aurre-prozesaketa bat egitea da, erregeletan oinarritutakoa. Itzulpen horren irteera SMT sistema batekin itzultzen da, zeinak azkeneko testua sortzen baitu [Ahsan et al., 2010]. Teknika honi esker sistema estatistikoak informazio gutxiagorekin ere lan egin dezake, eta erregelak ere sinpleagoak izan daitezke, gizakiaren lana nabarmen txikituz.

Itzulpen automatikorako metodo guzti hauek hala ere, zailtasunak dituzte, eta modu ezberdinean aurre egiteko ahalmena. Zailtasun hauen artean kokatzen dira, aurretik aipatutako anbiguotasuna, hizkera kolokial edo formala bereiztea, baita izen bereziak, entitate izenak, datak etab. identifikatzea ere.

Gaur egun, muga hauek direla eta, ez da inongo murriztapenik gabeko itzulpen automatiko perfekturik existitzen, nahiz eta programa asko dauden itzulpen erabilgarria egiten dutenak. Adibidez, oso ezaguna den *Google Translate*[1].

Era berean, mugak gainditu eta kalitate hobeko itzulpena egiteko helburua du TectoMT itzultzaile automatikoak. Sakoneko sintaxia erabiltzen du: informazio linguistikoa baliatuz, hizkuntzarekin ahal bezain lotura gutxi duen adierazpena sortzen da, batetik bestera pasatzea erraztuz.

Hala ere, sistema honek ere ez du oraindik behar bezalako itzulpenik egiten. Testuinguru honetan kokatzen da dokumentu honek aurkezten duen proiektua. Itzultzaile horretan hobekuntza bat gehitu nahian, rol semantikoak erabiliko dira. Azken hauek, esaldi bateko ekintzari buruzko informazioa ematen dute; ekintza nork, nori, noiz etab. egin dion zehazten da. Izan ere, ez du zertan subjektuak izan ekintzaile. Agian objektu bat izendatzen du, edo ekintza jasaten duena, eta aditzaren ondoren agertzen da egilea.

(2) a. *Yesterday, John hit Kevin with a hammer.*
b. *Yesterday Kevin was hit by John with a hammer.*
c. *Atzo, John-ek Kevin mailu batez jo zuen.*

Hizkuntza batean bertan, baita hizkuntzaz aldatzean ere, sintaxiak ezberdinak izan daitezkeela ageri da hiru adibideetan. Hala ere, hiru esaldietan ekintza jotzea da, John-ek burutzen du, Kevin-ek jasan, mailu bat erabili da horretarako eta kontatu baino egun bat lehenago gertatu zen.

[1] https://translate.google.es/

Hizkuntza batetik bestera pasaz rol hauek mantendu egiten direla ikusiz, itzulpenean oso baliagarria izan daitekeela pentsatu da. Horregatik, proiektu honen helburua, rol semantikoez baliatuz TectoMT tresnak egiten duen ingelesetik euskararako itzulpena hobetzen den ala ez ikustea izango da. Lan honetan, esaldietako rolak identifikatu eta etiketatzen dituzten tresnak aztertuko dira, horietako bat aukeratu, itzultzailean integratu, eta probak egin. Datu asko hartu beharko dira kontuan, itzultzailearen kalitatea, rolak etiketatzen dituzten tresnena etab. Hau guztiaren arabera, itzulpenean lagundu edo kalte egiten duen neurtu ahal izango da.

2. KAPITULUA

Proiektuaren Helburuen Dokumentua

Kapitulu honetan proiektuaren planifikazioa azaltzen da. Bertan proiektuaren irismena, denboraren plangintza, kalitatearen kudeaketa, komunikazio plana eta arriskuen kudeaketa aipatuko dira.

2.1 Proiektuaren irismena

Proiektu honen helburua, TectoMT itzultzaileko ingelesa-euskara bikotean, rol semantikoak automatikoki etiketatzen dituen tresna bat integratu eta itzulpenak hobera egiten duen ala ez aztertzea da.

Horretarako, ingelesezko etiketatzaile ezberdinak saiatuko dira, sisteman txertatzeko egokiena aukeratuz. Ondoren, tresna hori sisteman integratuko da, jatorrizko hizkuntzako esaldiak etiketatzeko. Behin hori lortuta, euskarara itzuli, eta euskarazko hitzei dagozkien ingelesezko hitzen etiketak eskuratu beharko dira. Ingelesezko rol semantikoei buruzko informazioari dagokion euskararako informazioa lortuta, helburuko esaldietan aplikatu daiteke.

Atal honetan, eginkizun hauek burutu ahal izateko beharrezkoak diren pausoak eta antolaketa azalduko dira: betekizunak zein diren, LDE diagrama, proiektuan zehar dauden emangarri eta mugarriak.

2.1.1 Betekizunak

TectoMT-ko lana proiektuaren oinarria izan arren, lan hori burura eramateko beste hainbat zeregin daude, eta hau bera ere atazatan banatu behar da. Ondotik ikus daitezke ataza ezberdinak:

1. Proiektuaren planifikazioa
   (a) Proiektuaren irismena
   (b) Ataza eta betekizunak zerrendatu
   (c) Denboraren estimazioa egin
   (d) Atazen egutegia zehaztu
   (e) Kalitate plana
   (f) Komunikazio plana
   (g) Arriskuen kudeaketa
      (i) Identifikazioa
      (ii) Ekiditeko plana
2. Ezagutzak eskuratu
   (a) TectoMT
   (b) Perl lengoaia
   (c) Rol semantikoak
   (d) Rol semantikoak automatikoki etiketatzeko tresnak
   (e) BVI lexikoia
3. Garapena
   (a) Rol semantiko etiketatzailearen aukeraketa
   (b) BVI-ren moldaketa
   (c) TectoMT-ko blokeen garapena
      (i) Etiketatzailea integratu
      (ii) Ingelesa-euskara informazioa lotu
      (iii) Euskaraz BVI-ko informazioa erabili
4. Probak
5. Kudeaketa
   (a) Jarraipen eta kontrola
   (b) Bilerak
6. Memoriaren garapena
7. Defentsa

2.1.2 LDE Diagrama

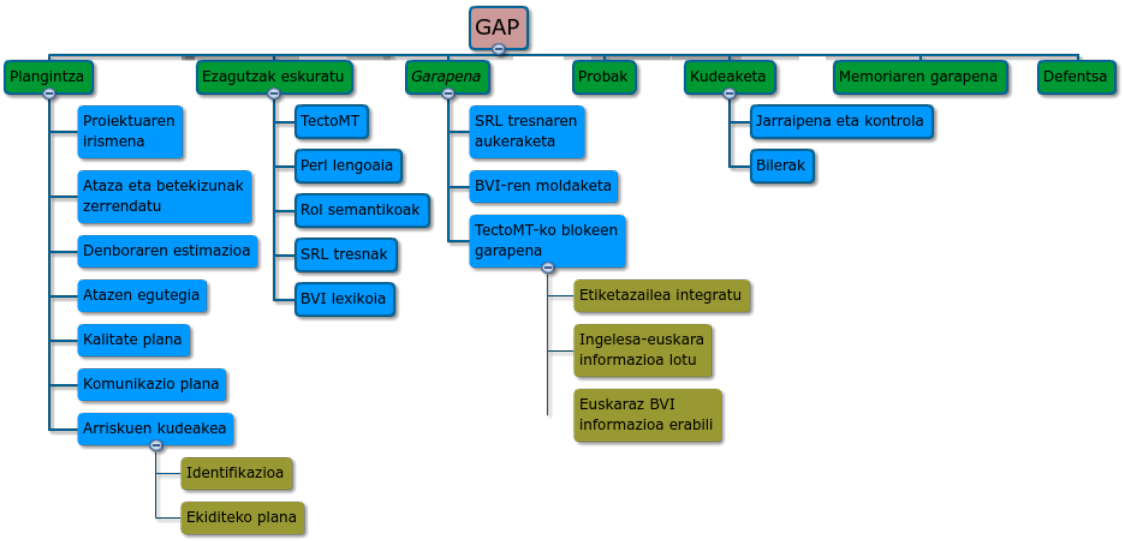
Ondoko 2.1 irudiko LDE diagramaren elementuen azalpenak hemen aurki daitezke:

- **GAP**: Gradu Amaierako Proiektua.

- **Proiektuaren planifikazioa**: Proiektuaren hasieran egin beharreko lana, proiektua nola burutuko den definitzeko.
  - Proiektuaren irismena: Proiektuaren helburua definitzen da hemen.
  - Ataza eta betekizunak zerrendatu.
  - Denboraren estimazioa egin: ataza bakoitza betetzeko zenbat denbora beharko den aurreikusi.
  - Atazen egutegia zehaztu: atazak burutzeko behar den denbora, batak besteekiko duen menpekotasuna eta lehentasuna kontuan hartuz, ataza guztiak denbora lerro batean kokatzea, proiektuaren hasieratik bukaeraraino.
  - Kalitate plana: Lana bera, memoria eta aurkezpenerako lortu nahi diren kalitate mailak definitzea.
  - Komunikazio plana: Proiektuko interesatuen artean komunikazioa nola egingo den definitzea.
  - Arriskuen kudeaketa: Proiektuan zehar gertatu daitezkeen ezustekoak identifikatzea, saihestu edo aurre egin ahal izateko modu bat aurkitzeko.
- **Ezagutzak eskuratu**
  - TectoMT: Itzultzaile sistema nola egituratua dagoen eta nola erabiltzen den ulertu behar da, ondoren bertan aldaketak egin ahal izateko.
  - Perl lengoaia: TectoMT, Perl lengoaian kodetua dagoenez hau ezagutzea ezinbestekoa da.
  - Rol semantikoak: Hizkuntzalaritzan behar diren oinarriak ezagutzea. Kasu honetan garrantzitsuena rol semantikoak zer diren ulertzea da, proiektuaren oinarria baita.
  - Rol semantikoak automatikoki etiketatzeko tresnak: Nolako tresnak kaleratuak dauden eta hauek zer egiten duten jakin behar da erabiltzen hasi aurretik.
  - BVI lexikoia: lexikoi bat zer den eta BVI lexikoiak zer nolako egitura duen ulertu behar da, proiektuan nola erabili daitekeen aztertzeko.
- **Garapena**: atal honetan etiketatzailearen aukeraketa eta kodeketa sartzen dira.
  - Rol semantiko etiketatzailearen aukeraketa: Mate eta ClearNLP tresnekin saiakerak egin eta aukera ezberdinak aztertu ondoren bietan bat aukeratuko da TectoMT-n erabiltzeko.

  - BVI-ren moldaketa: lexikoia aztertu eta behar diren aldaketak egin beharko dira, TectoMT-ko hitz berdinak erabiltzeko. Gainera BVI euskaratik ingelesera pasatzeko pentsatua dago eta proiektuan alderantziz erabili beharko da.
  - TectoMT-ko blokeen garapena: Rol etiketatzailea TectoMT-n integratu eta hauek kontuan hartzeko kodean egin beharreko aldaketak kokatzen dira hemen, konkretuki proiektuko helburua betetzea. Lan hau atal ezberdinetan banatuko da:
    * Etiketatzailea integratu
    * Ingelesa-euskara informazioa lotu
    * Euskaraz BVI-ko informazioa erabili

- **Probak**: Ataza hau iterazio moduan landuko da. Proben emaitzak aztertuz kodean beharrezko aldaketak egingo dira, pixkanaka emaitza ezberdinak lortuz.
- **Kudeaketa**: Proiektuaren aurrerapenaren kontrola.
  - Jarraipen eta kontrola: burututako ataza bakoitzaren kudeaketa, aurreikusitakoarekin konparaketa. Behar izan den denbora edo sortutako arazoak kontrolatzen dira.
  - Bilerak: zuzendariekin egindako bilerak dira hauek. Proiektuaren egoerari buruz edo sortu daitezkeen dudei buruz hitz egingo da.
- **Memoriaren garapena**
- **Defentsa**: Aurkezpena eta horretarako gardenkiak prestatzea.



2.1 Irudia: LDE diagrama

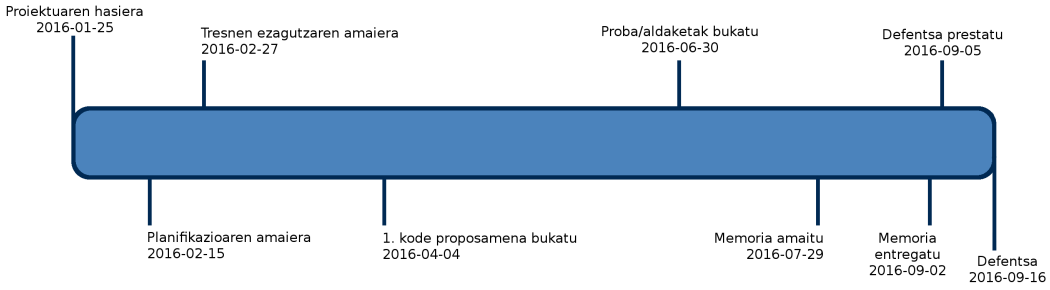
2.1.3 Emangarriak

Proiektu honetarako lau emangarri finkatu dira:

- TectoMT-ren bertsio berria: proiektuarekin lotuta egindako aldaketak, sortutako kodea.
- Proiektuaren garapenerako plangintza.
- Proiektuko memoria.
- Proiektuko aurkezpenerako gardenkiak.

2.1.4 Mugarriak

Ondoko irudian aurreikusitako mugarriak eta hauen epeak ageri dira. Epeak aste hasiera edo amaieren arabera jarri dira, memoria entregatzeko epeaz gain ez baitago kanpotik ezarritako beste mugarririk:



2.2 Irudia: Proiektuko mugarriak

2.2 Denboraren plangintza

Atal honetan proiektuaren atal edo ataza bakoitzeko beharko den denbora estimatuko da, horren araberako plangintza osatuz proiektu osorako.

2.2.1 Atazak eta hauen denbora estimazioa

LDE diagramako betekizun gehienak dira hemen aipatuko diren atazak:

1. Proiektuaren planifikazioa: 30 ordu.
2. Ezagutzak eskuratzea: 15 ordu.
   (a) TectoMT: 6 ordu. Zenbaki hau oso aldakorra da. Ezertan hasi baino lehen 6 ordu inguru behar direla estimatzen da oinarrizko ezagutzak eskuratzeko. Hala ere lanean hastearekin batera ezagutuko da sistema hobeto.
   (b) Perl lengoaia: 4 ordu. Ordu kopuru hau ere hasierako proba batzuk egiteko estimatzen da, lanean hastean ezagutzak sakontzen baitira, pixkanaka.
   (c) Rol semantikoak: 2 ordu.
   (d) SRL tresnak: 2 ordu.
   (e) BVI: 1 ordu.

3. Garapena: 98 ordu.
   (a) SRL tresnaren aukeraketa: 22 ordu.
   (b) BVI-ren moldaketa: 6 ordu.
   (c) TectoMT-ko blokeak: 70 ordu.
4. Probak: 130 ordu. Proben emaitzen arabera asko alda daiteke hau, hasieratik emaitza oso onak edo oso txarrak izan baitaitezke.
5. Kudeaketa: 25 ordu.
   (a) Jarraipena eta kontrola: 15 ordu.
   (b) Bilerak: 10 ordu.
6. Memoria: 90 ordu.
7. Defentsa: 20 ordu.

Orotara 408 ordu aurreikusi dira proiekturako. Proiektuak 12 kreditu dituenez, gutxienez 300 ordu sartzea aurreikusi behar da. Lan poltsa baten barne egingo denez ordea, ordu kopuru hori ezberdina da.

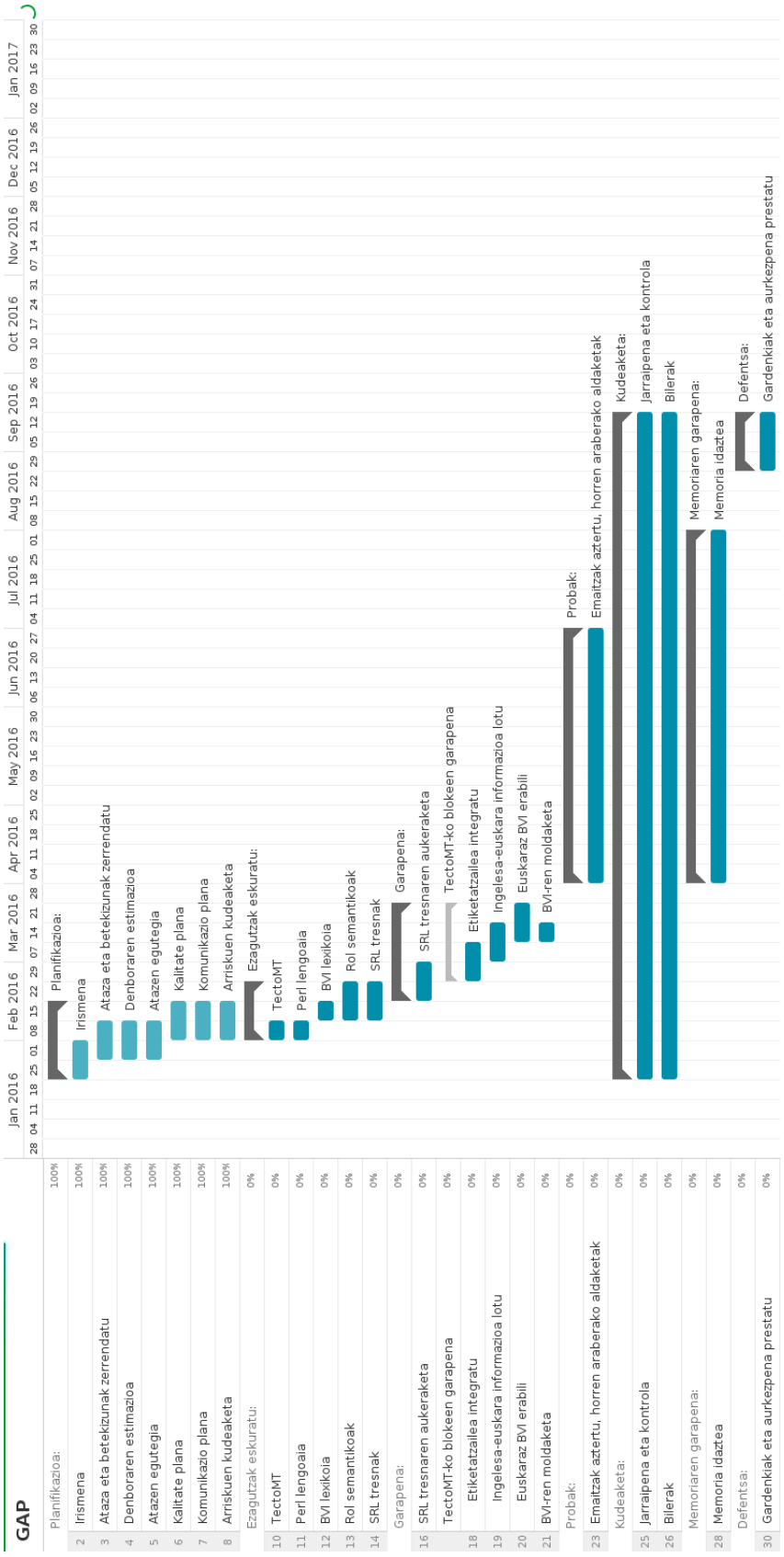
Esan beharra dago hala ere kopuru hau ez dela osoki egokia. Izan ere, Iazko urtean IXA taldean egindako praktiketan TectoMT sistema eta Perl lengoaiarekin lan egin zen, 450 orduko lan-poltsa batean. Bertan egindako lanek ezagutzak eskuratzeko balio izan du eta beraz proiektuaren parte dira. Otsailean, proiektuarekin hastean, TectoMT eta Perl ez dira hutsetik ezagutu beharko, berriz landu baizik.

2.2.2 Atazen planifikazioa

Ondoko 2.3 irudiko Gantt diagraman atazen planifikazioa ikus daiteke. Epeak finkatzerakoan ondoko puntuek eragina izan dezaketela pentsatu da:

- Unibertsitatea: Hautazko ikasgai batzuetako klase eta lanak izango dira proiektuaren hasieran, aste santura arte. Horregatik, nahiz eta proiektua hasieratik landu, memoria adibidez aste santutik aurrera gehiago garatuko da.
- Arrazoi pertsonalak: Unibertsitatetik kanpoko jarduerek eskatzen duten denbora kontuan hartu da.

IXA taldean garatuko denez proiektua, goizero bertan lan egin eta arratsaldetan klaseetara joatea aurreikusten da. Lanean proiektuarekin loturarik ez duten zereginak egin behar izanez gero ere, parte handiena proiektura zuzenduta egongo da.



2.3 Irudia: Proiektuko Gantt diagrama

2.3 Komunikazio plana

Atal honetan komunikazio plana aurkezten da. Bertan proiektuko interesatuak identifikatzen dira, baita hauekin komunikatzeko modu eta maiztasunak ere.

2.3.1 Interesatuen identifikazioa

- **Oneka Jauregi**: Proiektuaren egilea da, beraz, lehen interesatua. 12 kreditu lortu eta ikasketak amaitzeko beharrezkoa du proiektua. Horregatik garrantzitsua da beretzat lana txukun eta garaiz amaitzea.
- **Proiektu zuzendariak**: Arantza Diaz de Ilarraza eta Gorka Labaka, proiektu zuzendariak dira, Unibertsitateko irakasle eta IXA taldeko ikertzaileak. Egilearen proiektu zuzendariak izanik proiektua ondo amaitu eta gainditzea interesatzen zaie. Bestetik, IXA taldeko ikertzaile diren heinean, TectoMT itzultzailerako hobekuntza bat ekar dezakeen proiektu oro da interesgarria haientzat.
- **IXA taldea**: Proiektu hau IXA taldearen barne garatuko da. Zuzendarien gisara, IXA talderako garatzen den edozein proiektu da interesgarri, itzultzailean hobekuntzak ekar baititzake.

2.3.2 Informazio trukea

Interesatuen arteko informazio banaketa modu ezberdinetan egingo da. Alde batetik, Gorka Labakarekin eguneroko hartu emana izango da, bulego berean lan egingo baita. Honek egunean eguneko jarraipena egingo du ikaslea lagunduz.

Bestetik, bi zuzendariek eta ikasleak harremana emailez mantenduko dute bilerak finkatzeko. Bilera bakoitzean ondokoa noiz egin erabakiko da behar ezberdinen arabera.

2.4 Kalitatearen kudeaketa

Atal honetan proiektuaren kalitatea definituko da. Honen barne lortuko den emaitzaz gain, proiektuko beste arlo guztiak daude, komunikazioa, memoria eta defentsa bezala. Bakoitzerako lortu nahi den kalitate maila minimoa finkatuko da, hori lortzeko planarekin batera.

2.4.1 Kalitate maila planifikatu

- **Komunikazioa**: Interesatuen arteko komunikazioa mantendu egin behar da. Behar adina informazio eman behar zaie proiektuaren egoeraren berri izateko, zerbait gaizki eginez gero ahalik eta azkarren detektatu eta zuzentzeko.
- **Produktua**:
  - Kalitate maila minimoa: rol semantikoak kontuan hartuz itzulpena egiten lortu behar da arazorik gabe.
  - Kalitate maila egokia: probak eginez itzulpenak hobera egitea lortzeak proiektuan hobekuntza ekarriko luke.
- **Memoria**: EHU-k ezarritako formatua eta argibideak beteko ditu txostenak. Edukia ahalik eta ulergarriena izango da, irakurketa arinduz.
- **Defentsa**: Ez dira gardenki gehiegi egingo eta aurkezpena arintzeko saiakerak egingo dira.

2.4.2 Kalitatearen kontrola

- **Produktuan**: Proiektu zuendariak eguneroko jarraipena egingo du, honela dena espero bezala doala ziurtatuko da. Bestetik, emaitzen azterketa zuzena izateko hizkuntzalarien laguntza eskatuko da.
- **Memorian**: Zuzendariek emandako aholkuak kontuan hartuko dira hobekuntzak egiteko. Idazketa amaitzean osorik berrirakurri eta zuzenduko da.
- **Defentsa**: Gardenkiak aurkezpen denbora mugatura egokituko dira eta zuzendarien aholkuak jarraituko dira. Aurkezpena jende gehiagori egingo zaio defentsa egunaren aurretik, kanpoko iritzi eta aholkuak jasotzeko.

2.5 Arriskuen kudeaketa

Atal honetan proiektuan zehar gerta daitezkeen ezustekoak aurreikusi nahi izan dira, honela hauek ekiditeko plan bat sortuz, edo saihestezinak diren kasuan kontrolatzeko.

2.5.1 Informazio galera

- **Azalpena**: TectoMT-n egindako lana edo dokumentazio aldeko lana galdu egin daiteke.
- **Eragina**: Eragin handia izan dezake, lana berriz egin beharrak denbora galera suposatzen baitu.
- **Probabilitatea**: IXA taldeko zerbitzarietako informazioa galtzeko probabilitatea baxua da, bestea galtzea berriz altua.
- **Jatorria**: TectoMT-n egindako lana galtzeko arrazoia zerbitzarietako arazo bat litzateke. Beste datuak galtzeko arrazoia konputagailua edo dena delako biltegiratzeko sistema hondatzea izan daiteke. Bi kasuetan fitxategiak nahi gabe ezabatzea ere gerta daiteke.
- **Prebentzio plana**: Segurtasun kopiak egingo dira. IXA taldeko zerbitzarietako arazoa izanez gero, teknikariekin hitz egingo da, hauek jadanik sistema hau erabiltzen baitute. Bitartean denbora galtzea saihesteko, memoria aurreratuko da.
  Dokumentazioa *Google Drive*-n biltegiratuz galtzeko arriskua murrizten da. Bulegoko ordenagailuan, baita etxekoan ere, astean behin *Google Drive*-eko datuen kopia bat gordeko da.

2.5.2 Aurreikuspenak ez betetzea

- **Azalpena**: Atazak burutzeko zailtasunak edukitzea, epeak ez errespetatzea eta abar sartzen dira arrisku honetan.
- **Eragina**: Honek proiektuan eragin handia izan dezake, epeak errespetatu ezean ezingo baita lana deialdi horretan itzuli.
- **Probabilitatea**: altua.
- **Jatorria**: Emaitzak aztertzen ezin da jakin zenbat iterazio egin beharko diren. Horrez gain, Proiektuen Kudeaketan plangintzak egiten ikasi arren ez da oraindik esperientzia handirik hartu, beraz posible da gaizki planifikatu izana eta amaieran denbora gutxiegi izatea. Gainera, proiektuan zehar gaixotzea posible da, ezbehar bat gertatzea, edota klaseetan lan karga handia izatea.
- **Prebentzio plana**: Proiektuarekin ahal bezain goiz hasiko da. Egunero lanera joanez epeak errespetatzen direla ziurtatzea errazten da. Jarraipen eta kontrola ongi eramanez ez litzateke arazorik egon behar. Hala ere zerbait gertatuz gero, ordu gehiago lan egingo dira atzerapena berreskuratzeko.

3. KAPITULUA

Baliabideak eta aurrekariak

Kapitulu honetan proiektuan zehar erabiliko diren tresna eta kontzeptuak azaltzen dira. Horretarako, lau atal bereizi dira: Rol semantikoak, BVI lexikoia [Aldezabal et al., 2013], rol semantikoak automatikoki etiketatzeko *Mate* [3] eta *ClearNLP* [2] tresnak, eta TectoMT [Žabokrtský et al., 2008] itzultzaile automatikoa.

3.1 Rol semantikoak

Semantika, hitz eta esaldien esanahia aztertzen duen hizkuntzaren prozesamenduko arloa, arlo zailenetakoa da. Gaur egun, sintaxiarekin asko lotzen da eta honen egitura baliatzen da, analisi sintaktikoa egiterakoan informazio semantikoa integratuz.

Hizkuntzalaritzan, rol semantikoek, esaldi batean predikatuak[1] deskribatzen duen gertaerari (errealitatean gertatzen den edozer) buruzko informazioa zehazteko balio dute. Helburua gertakaria bera eta haren parte hartzaileen arteko erlazio semantikoak finkatzea da, ekintza nork nori egin dion, noiz, non etab. deskribatzea alegia.

Hainbat ikerketa eraman dira 60. hamarkadaz geroztik, rol horiek zein diren erabakitzeko. *Fillmore*-ren [Fillmore , 1968] kasu gramatikako lanek bultzatu dituzte ikerketa hauek. Denak ados daude rol semantikoen kontzeptu globalarekin; predikatu baten argumentuak orokortzea da, egitura sintaktikoetan semantika erregulartasunak identifikatzeko.

[1] Subjektu bati buruz baieztatzen dena da. Aditz batez osatua egoten da normalean, eta subjektuari buruzko beste atributuez. "Arantxak begi urdinak ditu"esatean, "begi urdinak ditu"predikatua da. Hala ere, bi predikatu mota daude: aditzezkoa eta izenezkoa.

Agente eta paziente bezalako rol nagusiekin ere ados daude, ekintza egin eta jasaten dutenak dira hauek. Hala ere, rol multzoak definitzerako orduan ezberdintasunak ageri dira ikertzaileen artean.

Ondoko 3.1 taulan rol multzo bat azaltzen da, nahiko orokorra dena.

Rolak eta azpirolak		**Definizioak**
Agentea	agent	Gertaera sortu duen pertsona
Instrumentua	instrument	Gertaera sortzen duen instrumentua/indarra
Tema/Pazientea	theme	Gertaerak eragiten duen objektua/pertsona
Esperimentatzailea	experiencer	Gertaerarekin psikologikoki edo fisikoki lotuta dagoen pertsona
Onuraduna	beneficiary	norentzat egiten den ekintza
Non/kokapena	location/at-loc	uneko kokapena
Noren/edukitzailea	possessor/at-poss	uneko edukitzailea
Zer-balioa	at-value	uneko balioa
Noiz	at-tme	uneko denbora
Nora/helburua	to-loc/destination	bukaerako kokapena
Norentzat/hartzaile	to-poss/recipient	bukaerako edukitzailea
Nora-balioa	to-value	bukaerako balioa
Nondik/jatorria	from-loc/source	hasierako kokapena
Norengandik	from-poss	hasierako edukitzailea
Nondik-balioa	from-value	hasierako balioa
Agentekidea	co-theme	ekintzaren bigarren tema
Bidea	path	zein bidetik zerbait pasatzen den

3.1 Taula: Rolen definizioak

(3) a. Anek liburua Joni saldu zion.
b. Liburua Aneren bidez saldu zitzaion Joni.

(3-a) esaldian Anek egiten du saltzearen ekintza, agentea da. Jon berriz pazientea izango da. (3-b) esaldian aurreko esaldiko ekintza eta parte hartzaile berdinak dira, esaldi egitura aldiz ez. Hala ere, rol semantikoek berdinak izaten jarraitzen dute.

Kasu hauetan agenteak (saltzailea), biziduna izan behar du eta saltzen dena (liburua) objektua izateko baldintza jarriko da. Baldintza hauei hautapen murriztapenak deitzen zaie.

Hona hemen rol semantikoen adibide gehiago:

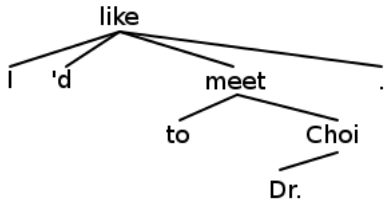
- Anek haragia labanaz moztu zuen. (Agentea + Instrumentua)
- Jonek leihotik sugea ikusi zuen. (Pazientea + Nondik/Jatorria)
- Liburua mahai gainean utzi nuen. (Tema)
- Jon beldurtuta zegoen. (Esperimentatzailea)

- Etxe bat eraiki zuen niretzat mendian. (Onuraduna + Kokapena)

Rol hauek esaldian identifikatzea zaila gertatu daiteke. Aipatu bezala, ez dagokio beti egitura gramatikal bera rol bakoitzari, adibideetan ageri da. Lan hori egiteko, automatikoki rolak etiketatzen dituzten tresnak sortu dira. Honela, esaldi bat emanda, lehenik dependentziak finkatzen dituzte. Hots, hitz bakoitzak besteekiko duen lotura, menpekotasuna. Esaldi batean aditz bat baino gehiago baldin badaude adibidez, seguruenik batzuk lehena eta besteak bigarrenarekin lotuta egongo dira.

Ondoko (4) esaldiko menpekotasunak ikus daitezke 3.1 irudian:

(4) *I'd like to meet Dr. Choi.*



ID	FORMA	BURUA
1	I	3
2	'd	3
3	like	0
4	to	5
5	meet	3
6	Dr.	7
7	Choi	5
8	.	3

3.1 Irudia: Dependentzia adibidea: zuhaitza eta taula

Goiko adibidean lehen zutabeak hitz bakoitzari dagokion identifikatzailea adierazten du, bigarrenak hitza bera eta hirugarrenak zein hitzen menpe dagoen, menpekotasunaren burua. *like* hitza esaterako ez dago ezeren menpe.

Dependentzia hauek markatu ondotik, esaldiko hitz bakoitzak zein paper jokatzen duen erabakitzen da, aditz nagusiari begiratuz. Hau guztiari SRL ere deitu ohi zaio (*Semantic Role Labelling*). Lan horretarako hainbat baliabide sortu dira eta euskararako BVI izena duen lexikoia garatu da azken urte hauetan. Hurrengo atalean azaltzen da zer den azken hau.

3.2 BVI - Basque Verb Index

Lexikoi bat hizkuntza edo ezagutza arlo bateko hitz zerrenda da. Hitz hauek lema edo lexema izan ohi dira, forma kanonikoko hitzak alegia. Berez, lexikoa da hitz batek edozein testuingurutarako berezkoa duen informazioa. Horregatik, ez du hitzaren informazio hutsa bakarrik, mota guztietako ezagutza barne har dezake: morfosintaktikoa, semantikoa... BVI-k gordetzen duen informazioa rol semantikoei buruzkoa da, euskararako.

Lexikoien artean badira ezagunenak diren batzuk, zabalenak ere bai, erreferentzia gisara erabiltzen direnak. Hemen *Verbnet* [Kipper, 2005] erabili da, gaur egun ingelesez dagoen aditz lexikoi zabalena. Egitura sintaktikoa, rol semantikoak (*agent*, *theme* etab.) edota hautapen murriztapenei (*+animate/-animate, +organization/ -organization, +location/-location* etab.) buruzko informazioa du.

PropBank [Palmer et al.,] berriz, fitxategi multzo bat da, non fitxategi bakoitzak aditz adiera bat edo gehiago baititu. Adiera bakoitzeko, argumentu zenbakitu bat eta aditzarekiko lotura handia duen rol semantiko zehatza ditu.

(5) Say: $\underset{\text{Arg0}}{\text{sayer}}$ (speaker), $\underset{\text{Arg1}}{\text{the thing said}}$ (utterance), $\underset{\text{Arg2}}{\text{recipient}}$ of the saying (hearer)

BVI lexikoia *EPEC-RolSem* corpusean oinarritzen da, eta azken hau EPEC, Euskararen Prozesamendurako Erreferentzia Corpusaren gainean eraiki da. EPEC corpusak 300 000 hitz ditu euskara estandarrean idatziak, morfologikoki eta sintaktikoki etiketatuta.

Helburua predikatu mailako informazioa gehitzea zen; horretarako sortu da *EPEC-RolSem* corpusa. *PropBank-VerbNet* eredua oinarri hartuta, roletako argumentuen informazioa etiketatua duen corpusa da. Beraz, *EPEC-RolSem* semantikoki etiketatutako EPEC corpusaren bertsioa dela esan daiteke.

Garapenean argiago azalduko da lexikoi honen erabilera, bi bertsio erabili baitira. 270 bat aditzeko zerrendaz osatua dago. Hauek, EPEC corpuseko 1457 aditz ezberdinen artean 30 agerpen edo gehiago dituztenak dira, *EPEC-RolSem*-en dauden berak.

BVI-ko informazioak ondoko egitura du, parentesien artean 3.3 taulako adibideari erreferentzia eginez:

- Euskarazko aditza
- Aditzaren adiera (*1 eta 2*)

- Adiera bakoitzari dagozkion ingelesezko aditz eta adierak (*_01 eta _02*)
- Argumentu bakoitza (*Arg0*), haren VerbNet-eko rol semantikoa (*agent*), EADB-ko (Euskal Aditzen Datu Basea) rol semantikoa (*esperimentatzailea*), deklinabide-kasua (*erg*) eta hautapen murriztapenak (*+biziduna/-biziduna, +gizakia/-gizakia, +konkretua/-konkretua...*).

AURKITU						
	1					
	find_01					
		Arg0	agent	esperimentatzailea	erg	+giz
		Arg1	theme	gaia	abs	
					par	
		Arg2	beneficiary	kokapena	ine	
					dat	
	2					
	feel_02/be_01					
		Arg0	agent	esperimentatzailea	erg	
		Arg1	theme	gaia	abs	
		Arg2	predicate	egoera	abs	
					ine	

3.3 Taula: BVI-ko informazioaren egitura

3.3 Rol semantiko etiketatzaileak

Rol semantikoen etiketatzea, edo SRL (*Semantic Role Labeling*), NLP-ko zeregin bat da. 3.1 atalean azaldutako rol semantikoak automatikoki identifikatzean datza.

Hainbat tresna daude eskuragarri zeregin hori egiteko. Lan honetarako, hauetako bi aukeratu dira. Proiektuaren garapen hasieran hauen arteko ezberdintasunak aztertu dira bietan bat aukeratzeko, emaitzen erabilerraztasunaren arabera. Hemen, bakoitzaren ezaugarriak azaltzen dira.

3.3.1 *Mate Tools*

Mate Tools-ek testua prozesatzeko lau tresna ditu: lematizatzailea, *part-of-speech* etiketatzailea (PoS, hitzaren kategoria: izena, aditza, preposizioa...), morfologia etiketatzailea (hitzen lema, erroa) eta dependentzia parserra. Ez du tokenizatzailerik[2], beraz sarrerako testua jadanik tokenizatua eman behar zaio.

Tresna guztiek *CoNLL 2009 Shared Task*-eko [1] formatua erabiltzen dute, token bakoitzarentzat gutxienez 12 zutabe tabulatu dituena. Token bakoitza lerro batean adierazten da, esaldiak lerro hutsaz banatuaz. Hauek dira zutabeek hartzen dituzten balioak, jakinez baliorik izan ezean azpimarra hartzen dutela:

- Id: uneko tokenaren identifikatzailea (esaldi bakoitzeko 1etik hasita).
- Form: hitzaren forma edo puntuazio ikurra.
- Lemma: hitzaren lema.
- PoS: part-of-speech etiketa, hizkuntzaren araberakoa.
- Feat: ordenatu gabeko ezaugarri morfologiko edota sintaktikoak ("|"-z bananduta). Adb. zenbakia, generoa, denbora etab.
- Head: uneko tokenaren buruaren Id-aren balioa, edo 0, ez badago ezeren menpe.
- Deprel: Buruarekiko duen menpekotasun harremana. Etiketak hizkuntzaren araberakoak dira. Head=0 bada, hemen seguruenik ROOT izango da balioa.
- Fillpred: Pred bete behar litzatekeen lerroetan Y balioa hartzen du.
- Pred: Predikatua; aditza eta adiera zehazten dira.
- Apreds: Predikatu bakoitzeko zutabe bat, esaldiko agerpen ordena berean, argumentu etiketekin.

[2]Tokenizazioa, testu sarrera bat token izeneko hitz edo sinboloetan banatzea da. Zuriune edo puntuazio ikurraz banatzen da token bakoitza

- PLemma, PPpos, PFeat, PHead, PDeprel: automatikoki aurresaten diren balioak dira.

(6) adibideko 3.4 taulan rol semantikoak etiketatu aurretiko dependentzia formatua ikus daiteke.

(6) *Those complaints don't appear to bother Apple.*

ID	FORM	LEMMA	PLEMMA	POS	PPOS	FEAT	PFEAT	HEAD	PHEAD	DEPREL	PDEPREL	FILLPRED	PRED	APREDs	
1	Those	those	those	DT	DT	_	_	2	2	NMOD	NMOD	_	_	_	_
2	**complaints**	complaint	complaint	**NNS**	NNS	_	_	3	3	**SBJ**	SBJ	_	_	_	_
3	**do**	do	do	**VBP**	VBP	_	_	0	0	**ROOT**	ROOT	_	_	_	_
4	n't	not	not	RB	RB	_	_	3	3	ADV	ADV	_	_	_	_
5	appear	appear	appear	VB	VB	_	_	3	3	VC	VC	_	_	_	_
6	to	to	to	TO	TO	_	_	5	5	OPRD	OPRD	_	_	_	_
7	bother	bother	bother	VB	VB	_	_	6	6	IM	IM	_	_	_	_
8	Apple	apple	apple	NNP	NNP	_	_	7	7	OBJ	OBJ	_	_	_	_
9	0	0	0	0	0	_	_	3	3	P	P	_	_	_	_

3.4 Taula: CoNLL 2009-ko formatuko dependentzia taula

Hemen, lehen lerroak taulako zutabe bakoitzaren izena adierazten du. Lerro bakoitzeko token bat eta bere informazioa ageri dira. Adibidean beltzez ikus daitekeen moduan, *do* aditza erro gisara definitu da, *complaints* izena haren subjektua izanik. FILLPRED, PRED eta APREDs-i dagozkien zutabeak oraindik bete gabe daude, rol semantikoekin lotutako eremuak baitira.

Erabilpena

*Mate Tools, Java*n inplementatuta dago eta *Java Runtime Environment* duen edozein makinatan exekuta daiteke.

Bi modutara erabiltzeko aukera dago: parseatze orokorra eginez edo rol semantikoen etiketatzea bakarrik gauzatuz.

1. Orokorra: Tokenizazioa salbu, beste etapa guztiak aplikatzen dira corpus oso bat parseatzeko. Sarrera fitxategia beraz, tokenizatua dagoela suposatzen du sistemak.

2. SRL bakarrik: Sarrerako fitxategiak *CoNLL 2009*ko datuen formatuan egon behar du, dependentzia zuhaitz egokiekin.

Exekuzioa

Webgunean [3] beharrezko fitxategi guztiak eskura daitezke. Hauetan *script*ak daude, parseatze osoa edo bakarra egiteko. Bertan, definitutako tokian, sarrera fitxategia, hizkuntza, hizkuntza horretarako entrenatutako ereduak eta irteera fitxategia zehaztu behar dira. Ondoren deia *script*ean bertan egiten da. Hala ere, *script*ean idazten den *java* bertsioa eta beharrezko paketeen bertsioak ordenagailuak dituenekin bat datozela egiaztatu behar da. Parseatzeko bi moduen arteko ezberdintasuna ereduetan egongo da, baita dei horretan ere.

Amaieran, itxura honetako taula bat lortzen da esaldi bakoitzeko:

1	Those	_	those	_	DT	_	_	_	2	_	NMOD	_	_		_	_
2	complaints	_	complaint	_	NNS	_	_	_	3	_	SBJ	_	_		A1	A0
3	do	_	do	_	VBP	_	_	_	0	_	ROOT	_	_		_	_
4	n't	_	not	_	RB	_	_	_	3	_	ADV	_	_		AM-NEG	_
5	appear	_	appear	_	VB	_	_	_	3	_	VC	Y	appear.02		_	_
6	to	_	to	_	TO	_	_	_	5	_	OPRD	_	_		C-A1	_
7	bother	_	bother	_	VB	_	_	_	6	_	IM	Y	bother.01		_	_
8	Apple	_	apple	_	NNP	_	_	_	7	_	OBJ	_	_		_	A1
9	.	_	.	_	.	_	_	_	3	_	P	_	_		_	_

3.5 Taula: CoNLL 2009 formatuko SRL irteera adibidea

Goiko 3.4 taularen antzekoa da, baina kasu honetan azkeneko lau zutabeak beteta daude. Y eremua duten tokenak dira predikatuei dagozkienak. Ondoko zutabean aditz adiera zehazten da eta azken bietan aditz bakoitzarekin lotutako argumentuak. Hala, azkenaurreko zutabean *appear.02*-ren argumentuak kokatzen dira. *appear* aditzerako, *complaints* A1 argumentua izango da, *n't* AM-NEG eta *to*, C-A1 argumentua. *bother* aditzerako berriz, *complaints* eta *Apple* A0 eta A1 argumentuak izango dira, hurrenez hurren.

3.3.2 *ClearNLP* paketea

ClearNLP proiektutik *ClearNLP* tresnak sortu ziren, zeintzuek zeregin ezberdinak egitea ahalbidetzen baitute: dependentzia zuhaitzetik zutabekako dependentzia formatura pasatzea, tokenizatzea, *part-of-speech* etiketatzea, analisi morfologikoa, dependentzia parseatzea eta rol semantiko etiketatzea.

Etiketatzailea martxan jarri aurretik tresna batzuk behar dira. *Mate* tresnako antzekoak dira, baina banatuta aurkezten dira, beraz, erabiltzaileak banan banan eskuratu behar ditu.

Honakoak dira:

- Hiztegi bat: adjektibo, aditz...zerrendak, onartu beharreko hitz zerrendekin.
- *Part-of-Speech* etiketatzeko eredu entrenatu bat
- Dependentziak parseatzeko eredu entrenatu bat
- SRL eredu entrenatu bat

Sarrera moduan fitxategi soil bat edo direktorio bat jarri daiteke.

ClearNLP-ren web orrian [2] behar den guztia instalatzeko lotura eta azalpenak daude, formatu ezberdinetarako adibideekin.

Sarrerako testuak ez du formatu berezirik behar. Irteera berriz ondoko zutabeetan egituratzen da, baliorik gabeko eremuetan azpimarra idatziz:

- Id: uneko tokenaren identifikatzailea (esaldi bakoitzeko 1etik hasita).
- Form: hitzaren forma.
- Lemma: hitzaren lema.
- PoS: *part-of-speech* etiketa.
- Feats: ezaugarriak (ezaugarri ezberdinak "|"-z banatzen dira eta gako eta balioak "="-ez).
- Head: menpekotasuneko buru den tokenaren identifikatzailea.
- Deprel: dependentzia etiketa.
- Sheads: 'semantic heads', etiketatutako argumentu eta rolak.

(7) *Those complaints don't appear to bother Apple.*

Irteera:

ID	FORM	LEMMA	POS	FEATS	HEAD	DEPREL	SHEADS
1	Those	those	DT	_	2	det	_
2	complaints	complaint	NNS	_	5	nsubj	7:A0;5:A1=PPT
3	do	do	VBP	_	5	aux	_
4	n't	not	RB	_	5	neg	5:AM-NEG
5	appear	appear	VB	pb=appear.02	0	root	_
6	to	to	TO	_	7	aux	_
7	bother	bother	VB	pb=bother.01	5	xcomp	5:C-A1
8	Apple	apple	NNP	_	7	dobj	7:A1=PPT
9	0	0	0	_	5	punct	_

3.6 Taula: *ClearNLP*-ko SRL irteera adibidea

Azken (7) adibideari dagokion 3.6 taulan dependentzia harremanak, predikatu eta argumentuak ageri dira. Taularen irakurketa nola egin azalduko da hemen, datu bakoitzaren

esanahi zehatza ez. *appear* 5. hitzari dagozkion argumentuak *complaints* (A1), *n't* (AM-NEG) eta *bother* (C-A1) dira. *bother* aditzarenak berriz *complaints* (A0) eta *Apple* (A1). Datu hauek azken zutabean irakur daitezke.

TectoMT-n, ingelesezko rol semantikoak etiketatzeko tresna hauetako bat erabiliko da. Ondoren, horiei dagozkien euskarazko rolen egitura eskuratu behar da. Horretarako, baliokidetzak zehazten dituen lexikoi bat erabiliko da. Ondoko atalean, TectoMT zer den argituko da.

3.4 TectoMT itzulpen sistema

TectoMT itzulpen automatikorako sistema da, Treex [5] NLP *framework*ean inplementatua, Pragako Charles Unibertsitatean.

Treex, hasieran TectoMT deitzen zena, Perl lengoaian inplementatutako modulu askotako NLP *software* sistema da. Itzulpenaz gain, beste hainbat NLP zereginetarako *software* irtenbide bat garatzeko laguntza da, moduluak berrerabilgarriak baitira, eta erraz moldagarriak.

Modulu batzuk bakarrik aplikatuz gero hainbat emaitza lortu daitezke, adibidez lematizatzaile bat, PoS etiketatzaile bat etab. Horregatik, TectoMT-ri izena aldatzea erabaki zen. Tresna multzoa Treex izendatu zen eta itzultzailea bera, tresna guztiak erabiltzea alegia, TectoMT. Beraz, azken finean, TectoMT edo Treex aipatuta ere, tresna berberez hitz egiten da.

Sistema honetan modularitateak garrantzia berezia du: itzulpena, modulu askoren sekuentzia gisara inplementatuta dago, blokeak deituak. Modulu bakoitzak ongi definitutako zereginak ditu, besteekiko independenteak. Hori dela eta, bakoitza bere aldetik hobetu ala ordeztu daiteke.

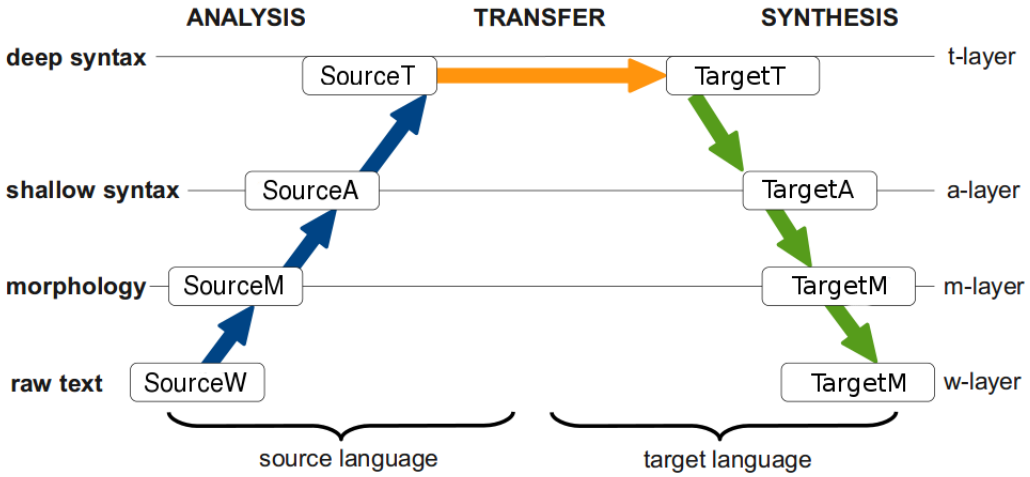
Tresna honek sakoneko sintaxia erabiltzen du. Esaldietan hitz bakoitzaren informazio zehatza etiketatzen da, ahal bezainbat. Adibidez, aditz baten denbora eta pertsona gorde daitezke. Informazio horiek edukita, aditza jokatu eta esaldian behar duen forma ematea posible da.

Informazio hauek hizkuntzarekiko ia independenteak izan daitezke. Honela, datu gramatikalak edukiko ditugu alde batetik, esaldiaren egitura, geruza bakoitzean gehitutako informazioa, eta bestetik hitza bera, lema.

Egitura sakon hori hizkuntza batetik bestera pasatzea ez litzateke lan handia izan behar. Adibidez, aditzaren denbora eta pertsona berdinak izango dira, baina hizkuntza bakoitzak ez ditu datu berdinak behar aditz bat jokatu ahal izateko. Aldaketa horiek egin beharko dira helburuko hizkuntzako egitura sakona lortzeko. Beraz, itzulpena egiteko, hitzak berak itzuli beharko lirateke alde batetik, lemak alegia, eta egitura sakona bestetik, aipatu bezala.

Erregeletan oinarritutako sistema gehienetan bezala, TectoMT-n ere transferentzia bidezko itzulpena erabiltzen da, hiru pausotan banatzen dena: i) jatorrizko hizkuntzako testuaren analisia, haren egitura gramatikala zehazteko; ii) transferentzia, lortutako egituratik

itzultzeko egokia den egitura batera, eta iii) helburuko testuaren sorkuntza. Ondoko 3.2 irudian 3 faseak ikus daitezke:



3.2 Irudia: TectoMT-ko itzulpen faseak

Analisian testua anotatu egiten da informazioa gordetzeko, eta sakoneko sintaxiako egitura lortzen da, hizkuntzarekiko ahalik eta independenteena.

Egitura hori lortuta, jatorrizko hizkuntzatik helburuko hizkuntzako egitura baliokidera pasa behar da: transferentzia. Aipatu bezala, ez litzateke aldaketa handirik egin behar.

Ondoren, helburuko hizkuntzako egitura sakonetik, helburuko esaldia sortu behar da: sorkuntza. Hemen, analisian egiten diren pausoen alderantzizkoa egiten da.

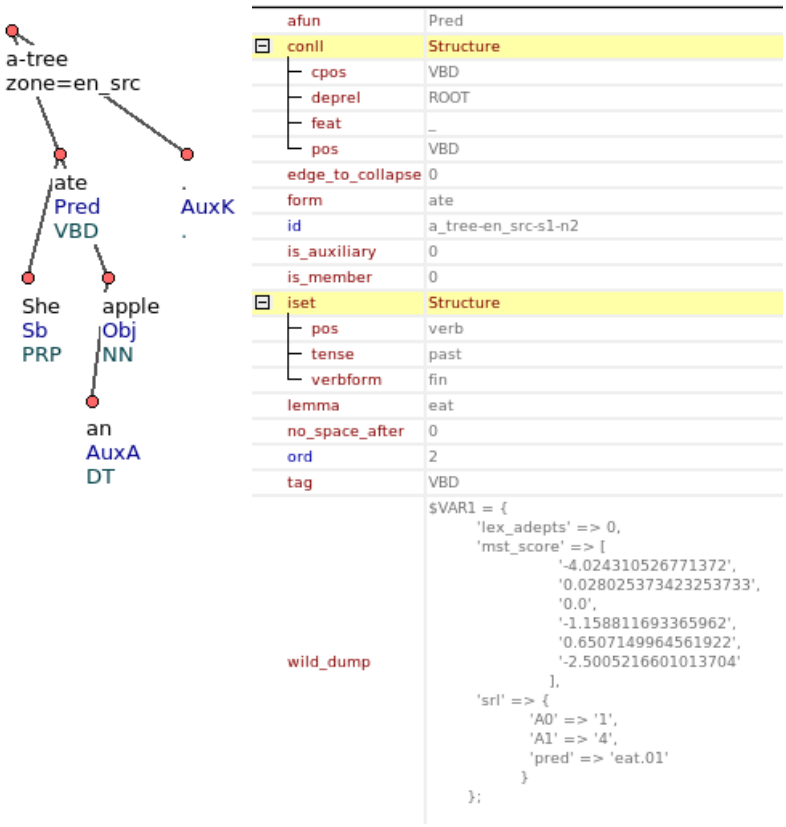
Hiru etapa hauetan zehar, irudian ageri den moduan, lau geruza definitzen dira, abstrakzio maila gorakorrean: testu gordina (*raw text*), geruza morfologikoa, azaleko-sintaxi geruza (geruza analitikoa), eta sakoneko sintaxi geruza (tektogramatikala). Geruza antolaketa hau *Prague Dependency Treebank*-en [Čmejrek et al., 2004] oinarritzen da (PDT). PDT txekieraz idatzitako testu bilduma handia da, 3 geruzatan egituratutako anotazio morfologiko, sintaktiko eta semantiko konplexuak dituena.

Geruza hauek aipatzeko w-, m-, a-, eta t- aurrizkiak erabiltzen dira, hurrenez hurren, bakoitzaren ingelesezko hitzaren lehen hizkitik hartuta.

- Testu gordina (w-geruza, *w-layer*): inongo anotazio linguistikorik gabekoa.

- Geruza morfologikoa (m-geruza, *m-layer*):
  Esaldi bakoitza tokenizatzen da, token bakoitza lema eta etiketa morfologiko batekin notatzen da. Honela, hitz bakoitzaren *part-of-speech* (izena den, aditza etab.) eta azpi-kategoria (zenbakia, generoa, denbora etab.) biltzen dira, hitzaren lemarekin batera (hitzaren forma kanonikoa, hiztegiko forma, Adb: “dokumentuaren” hitzerako, “dokumentu”).

- Geruza analitikoa (a-geruza, *a-layer*):
  Esaldi bakoitza azaleko-sintaxi dependentzia-zuhaitz baten bidez irudikatzen da (a-zuhaitz, *a-tree*) eta a-nodo (*a-node*) egitura definituta dago. Bertan, m-geruzako token bakoitzari a-geruzako nodo bat dagokio. a-nodo bakoitzak anotazioa du, bere nodo gurasoarekin duen dependentzia harremana adieraziz. Ondoko 3.3 irudian a-zuhaitz bat ikus daiteke, *ate* a-nodoari dagozkion datuekin.



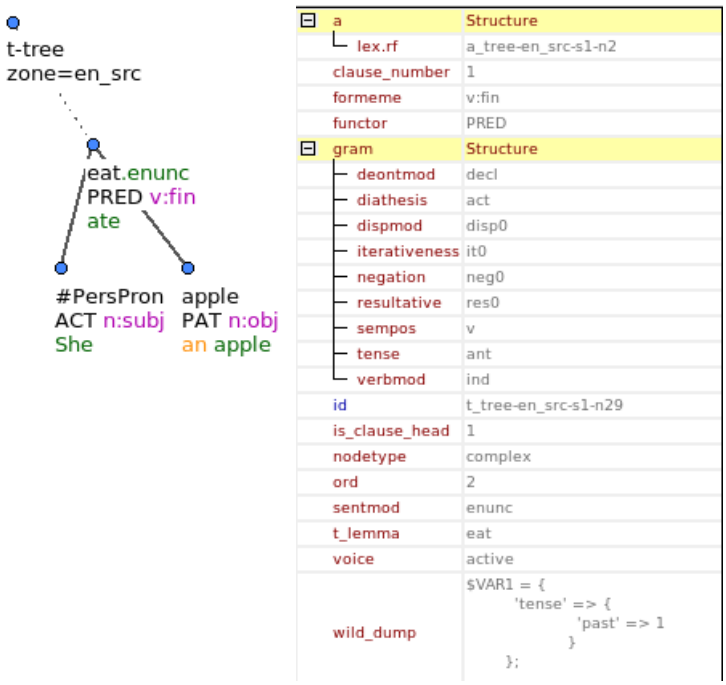
3.3 Irudia: a-zuhaitza eta a-nodoen informazioa

- Geruza tektogramatikala (t-geruza, *t-layer*):
  Esaldi bakoitza sakoneko sintaxia dependentzia-zuhaitz gisara irudikatzen da (t-zuhaitz, *t-tree*), hemen ere zuhaitzean t-nodo (*t-node*) egitura izanik. Informazio

zehatza darama zuhaitzak, hizkuntzarekiko beste geruzetan baino independenteagoa. Nodoak hitz autosemantikoak (esanguratsuak) dira. Hitz hauek garraiatutako informazioa, aditz laguntzaile, preposizio etab. bezala, t-nodoen atributu bidez erakusten da. Hona atributu garrantzitsuenak:

- Lema tektogramatikala: aurreko geruzetako lemaren antzekoa
- *Grammateme*-ak: kategoria morfologikoen baliokideak, adb. denbora, zenbakia.
- *Formeme*-ak: azaleko geruzan erabiliko den forma morfosintaktikoa gordetzeko. n:subj formemak izena subjektuaren posizioan errepresentatzen du, n:abs+X-k izena kasu absolutiboan eta v:ger -ek aditza gerundioan, besteak beste.

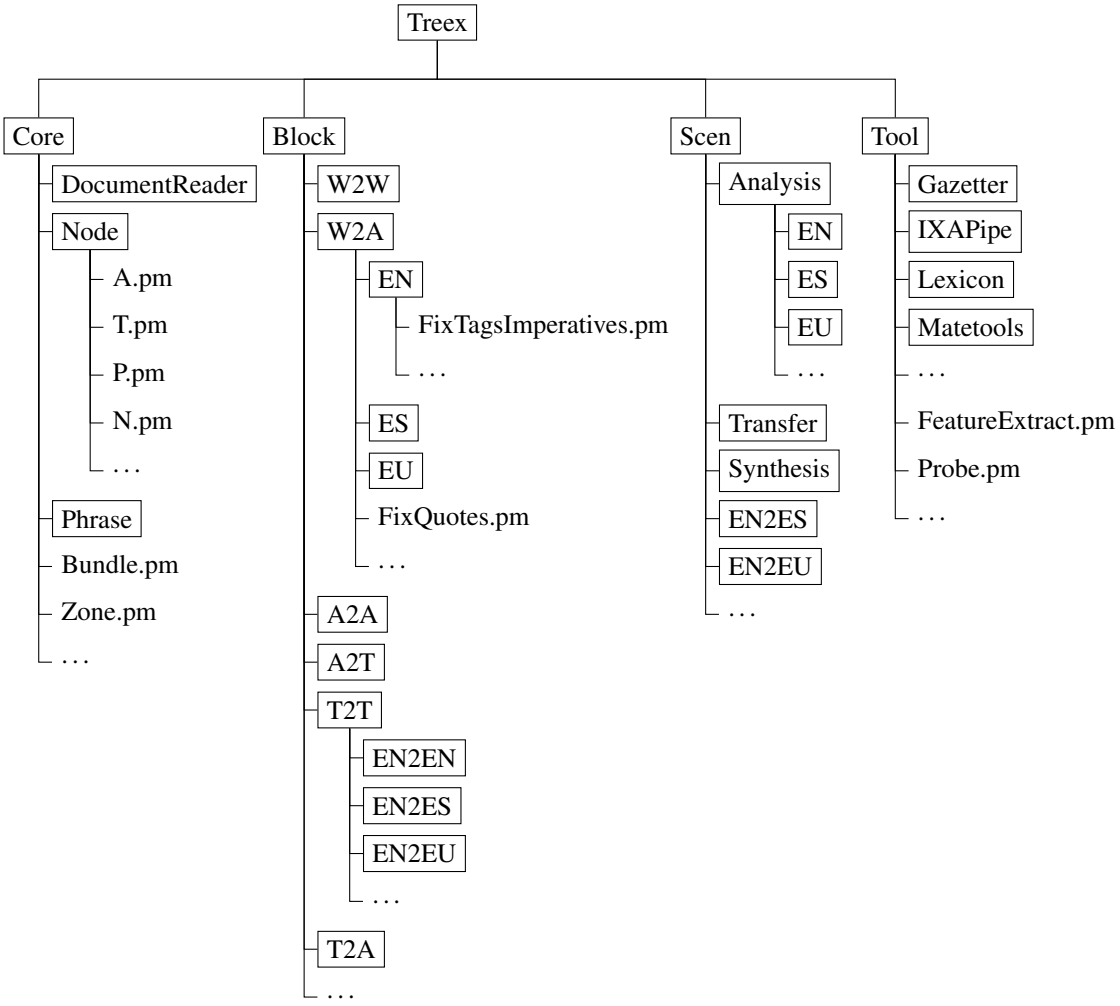
Ondoko 3.4 irudian t-zuhaitz adibide bat ageri da, *eat* hitzari dagozkion informazioekin. Geruza analitikoko esaldi bera aztertu da hemen. 3.3 iruditik hona, ikus daiteke nola *ate* aditzaren forma orokortu eta *eat* den orain. Gainera, *an* preposizioaren nodoa desagertu egiten da, *apple* nodoaren atributu bilakatzeko. Esan bezala, esanguratsuak diren hitzak dira nodoa mantentzen dutenak.



3.4 Irudia: t-zuhaitza eta t-nodoen informazioa

3.4.1 TectoMT-ren egitura

Sistema multzo ezberdinetan banatzen da. Ondoko diagraman 4 multzo nagusiak ageri dira. Ondoko ataletan bakoitza azalduko da.



3.4.1.1 Nukleoa (*Core*)

Hemen Treex-en egitura definitzen da: dokumentua, esaldiak, nodoak, zuhaitzak, *bundle* eta *zone* egiturak.

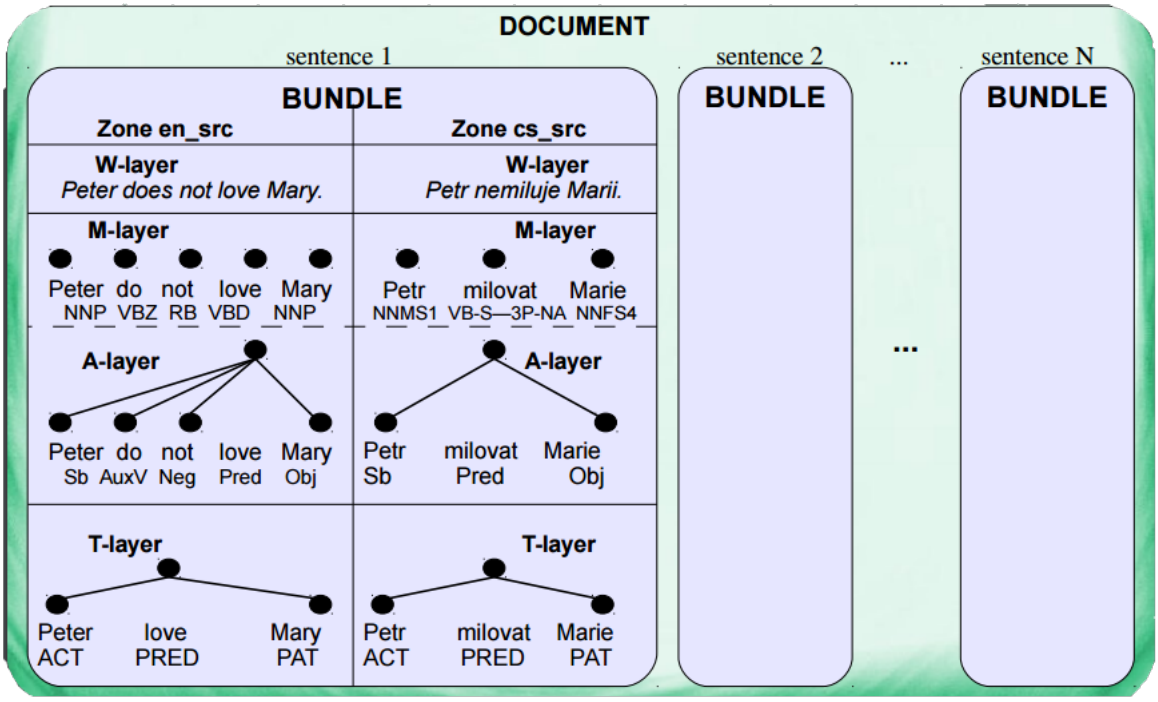
Aurretik aipatutako a-geruza eta t-geruzez gain, N eta P geruzak ere definitzen dira. N-geruza entitate izenei dagokie eta p-geruzan berriz esaldi-egitura zuhaitzak gordetzen dira. m-geruzako informazioa a-geruzan gordetzen da.

Dokumentu batean aztertu beharreko esaldi guztiak daude. Esaldi bakoitzeko *bundle* eremu bat sortzen da. Bertan, esaldi horretarako dauden bertsioak gordetzen dira, zuhaitz

eta hizkuntza ezberdinetakoak. Honen helburua informazio guztia gordetzea da, transferentzia egiterakoan analisiko datuak ez galtzeko.

Bundle bakoitzeko eremu (*zone*) ezberdinak finkatzen dira. Hizkuntza bakoitzeko eremu bat sortzen da, baita hautatutako parametro ezberdinen arabera ere. Adibidez, eremu batek jatorriko edo helburuko esaldia den adieraz dezake. Egitura honen barruan bi hizkuntzetako zuhaitzak gordetzen dira.

Ondoko 3.5 irudiak egitura hau argi erakusten du. Esaldi baten ingelesezko eta txekierazko bertsioak biltzen dira, biak jatorrizko hizkuntzan, hau da, batetik besterako itzulpena izan gabe.



3.5 Irudia: TectoMT-ko esaldien egituraketa

3.4.1.2 Blokeak (*Block*)

Block multzoan aurretik aipatutako blokeak aurkitzen dira, Perlen kodetuta. Bloke bakoitza fitxategi bat da, eta bertan agindu zehatzak daude, zuhaitzei eragiten dietenak. Batzuk hizkuntza bakarrari aplikatzeko dira, besteak biei batera, eta badira hizkuntza batekiko dependentziarik ez dutenak ere. Batzuk geruza batetik mugitu gabe egiten dira, eta besteak geruzaz aldatzeko behar dira. Fitxategiak direktorio ezberdinetan antolatuta daude horren arabera.

Bloke bakoitza erraz moldagarria da, baita ordezkagarria ere. Hizkuntza ezberdinetarako aplikagarriak direnez, hizkuntza bikote baterako garatutako blokeak beste bikote baterako erabil daitezke. Ingelesetik euskararako itzulpenean adibidez, ingelesa jatorri duen beste bikote bateko blokeak erabil daitezke, edo euskara helburu duen batekoak.

Block direktorioko antolaketa ongi irudikatzeko adibide bat erabiliko da:

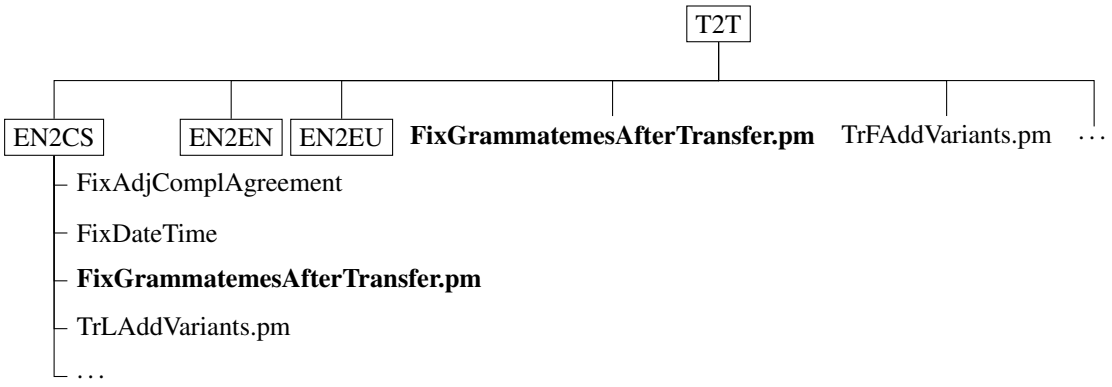
- A2A (A to A, ingelesez), bertan aurki daitekeen direktorio bat da. A geruzatik A geruzarako aldaketak eragiten dituzten blokeak gordetzen ditu, hau da, geruza aldatzen ez dutenak. Beraz, hizkuntza ere ez. Jatorrizko hizkuntzarentzat zein helburukoarentzat izan daiteke blokea.

  Direktorio horretan bloke batzuk daude, ahal bezain izen esanguratsuekin: AddDirectObjectMarkers.pm, AddPluralMarkers.pm, CopyAtree.pm ... Fitxategi hauek hizkuntza guztietarako erabili daitezkeenak dira. Batzuk hizkuntzarekiko independenteak izango dira, eta beste batzuk ez osoki.

  Bertan beste direktorio batzuk ere badaude, hizkuntzaka: EN, ES, EU... Adibidez, A2A/EN/ direktorioan honako fitxategiak aurki daitezke: MateSRL.pm, Retokenize.pm... Lehenengo hau esaterako proiekturako sortu da.

- T2T direktorioa ezberdina da, hizkuntza batetik besterako transferentzia geruza honetan egiten baita. Hala ere, aurreko azalpenaren antzera, hemengo blokeetan ere posible da hizkuntza bakarrean aldaketak eragitea, T geruzan geldituz. Horregatik, honela egituratzen da direktorio hau:

  - T2T/EN2EN: hau aurreko azalpeneko A2A/EN -en baliokidea litzateke; T geruzan, ingelesezko testuari aldaketak eragiten dizkioten blokeak daude hemen, jatorrizko zein helburuko hizkuntza izanik.
  - T2T/EN2EU: honelako direktorioetan hizkuntza aldaketa egiten duten blokeak daude. Kasu honek ingelesetik euskararakoa adierazten du, EU2EN ere egongo da.



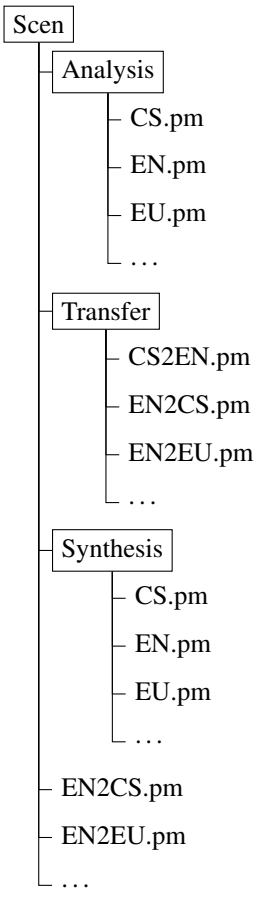
Blokeetan, objektuetara orientatutako egitura erabiltzen denez, herentzia ere aplikatzen da. Bloke orokor batzuk daude, hizkuntzekiko independenteak direnak. Hizkuntza bakoitzeko berezitasunen arabera funtzioak birdefini daitezke.

3.4.1.3 Eszenarioak (*Scen*)

Hemen eszenario ezberdinak definitzen dira. Hauek bloke segidak dira. Aipatu bezala, blokeak berrerabilgarriak dira eta bata bestearekiko independenteak. Itzulpena edo beste NLP zeregin bat burutzeko ez dira bloke berak behar. Beraz, Treex-ek moduluak nahi den gisara konbinatzea ahalbidetzen du.

Eszenarioak fitxategiak dira. Bertan, blokeak zerrendatzen dira, nahi den ordenan, beharrezko parametroekin. Itzulpenerako esaterako, zein bloke exekutatu nahi den finkatzen da, analisirako, transferentziarako eta sorkuntzarako.

Scen direktorioa honela egituratzen da:



Bertan dauden EN2ES.pm, EN2EU.pm... moduko fitxategiek ez dute zuzenean bloke zerrenda bat, beste eszenario batzuetarako deia baizik. Esaterako, EN2EU.pm fitxategitik Analysis/EN.pm, Transfer/EN2EU.pm eta Synthesis/EU.pm fitxategiei dei egiten zaie.

Honela, hizkuntza bikoteetarako konbinaketa ezberdinak egiten dira, eszenarioak errepikatu beharrik izan gabe. Izan ere, ingelesetik euskarara edo txekierara itzulpena egitean, ingelesaren analisia bera izango da.

Analisi, transferentzia eta sorkuntzako fitxategiak azaldutako eszenarioak bezala antolatuta daude. Hizkuntza bakoitzerako eszenarioak finkatzen dira eta fitxategi barruan, exekutatu beharreko bloke guztien izenak agertzen dira, erabakitako ordenan. Hona ingeleserako analisirako adibide baten zati bat:

```
my $scen = join "\n",
    'W2A::EN::Tokenize',
    'W2A::EN::NormalizeForms',
    'W2A::EN::FixTokenization',
    $self->tagger eq 'Morce' ? 'W2A::EN::TagMorce' : (),
    $self->tagger eq 'MorphoDiTa' ? 'W2A::EN::TagMorphoDiTa' : (),
    'W2A::EN::FixTagsImperatives',
    'W2A::EN::Lemmatize',
    'A2N::EN::DistinguishPersonalNames',
    'W2A::EN::ParseMST model=conll_mcd_order2_0.01.model',
    'A2T::EN::FixEitherOr',
    'A2T::EN::SetFunctors',
    'A2T::MarkParentheses';
```

3.4.1.4 Kanpoko tresnak (*Tool*)

Direktorio hau kanpoko tresnekin lotura egiteko erabiltzen da. Bloke orokor bat da, zeinetan edozein kode jarri daitekeen. Tresna hauek, TectoMT-tik kanpokoak izan ohi dira: IXAPipe adibidez, edo java erabiltzen dutenak. Direktorio honetan kokatzen da Mate Tools, proiektuan rol semantikoak etiketatzeko erabiliko dena.

Tresna bakoitza Treex-ek erabiltzeko moduan jarri behar da. Lengoaia ezberdinen arteko lotura egin eta honi deia egiteko beraz modulu bat sortzen da Tool direktorioan. Ondoren blokeetatik zuzenean erabiltzeko aukera emango du honek. Dei hori burutzeko, hau da, tresna erabiltzeko deskargatuta eduki behar diren liburutegi, JAR fitxategi etab., *share/installed_tools* kokapenean daude.

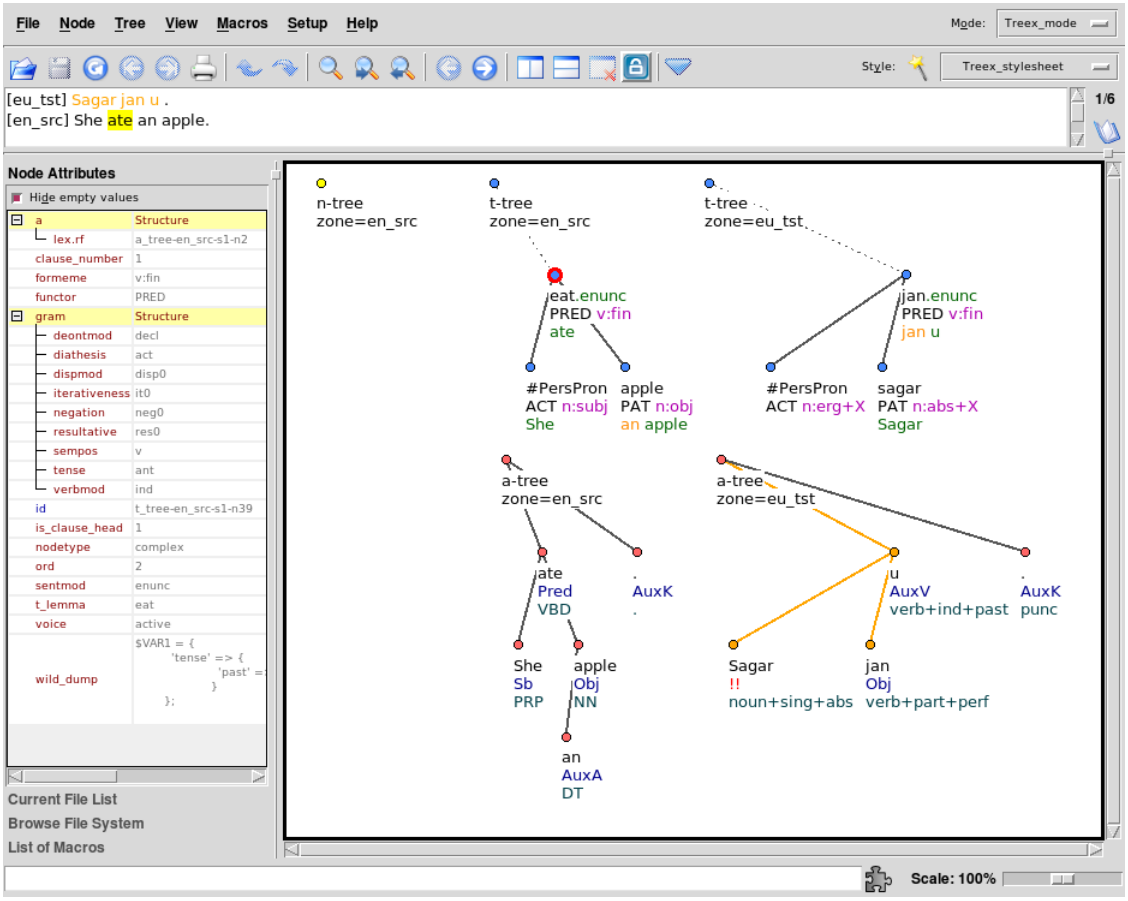
Horrez gain, tresna ongi integratu dela egiaztatzeko, beste modulu bat ere sortzen da. Bertan eskuz adibide zehatz batzuk sartzen dira, emaitza zein izan beharko litzatekeen ezagutuz. Honela erabilera egokia ziurtatu daiteke modu azkarrean.

3.4.2 Emaitzak

Behin TectoMT-ren egitura ulertuta, honen erabilera azalduko da, itzulpenak egiteko. Jatorriko eta helburuko hizkuntzak, itzuli nahi den testua eta eszenarioa finkatu ondoren, itzulpena martxan jartzen da. Horrez gain, helburu hizkuntzako erreferentzia testu bat ere definitu daiteke. Esaldiz esaldi, bloke guztiak exekutatzen dira.

Emaitzak modu ezberdinetan gordetzen dira. Alde batetik, sarrera fitxategiaren moduko esaldi zerrenda bat sortzen da, baita jatorri-helburu esaldi bikoteka ere. Sarreran helburu hizkuntzako erreferentzia testuren bat pasaz gero, hori ere hemen agertzen da. Honela, konparaketak egin daitezke bai eskuz, edo makinaz ere. Erreferentziazko testua sortutakoarekin konparatu eta ezberdintasunak nabarmendu daitezke. Azkenik, esaldi bakoitzeko fitxategi konprimitu bana sortzen da. Hauek TrEd (*Tree Editor*) interfazean bistaratzeko baliagarriak dira. Bertan, esaldiaz gain sortutako zuhaitz eta informazio guztia dago.

TrEd interfazean honelako itxurako leihoak ikus daitezke, zuhaitz eta datu guztiekin. Hauek eskuz moldagarriak dira, bistaratzea errazagoa izan dadin:



3.6 Irudia: TrEd interfaze adibidea

Bestetik, itzulpenaren kalitatea zenbakiz neurtzeko tresna bat ere badago: BLEU zenbakia.

BLEU (*Bilingual evaluation understudy*) [Papineni et al., 2002] hizkuntza batetik bestera automatikoki itzulia izan den testuaren kalitatea neurtzeko algoritmoa da, gaur egun erabilienetakoa. BLEU-k oinarritzat duen kalitatea, pertsona batek egindako itzulpena eta makinak sortutakoaren arteko antzekotasunean datza.

Puntuazioa kalkulatzeko, testua zati indibidualetan tratatzen da, esaldietan normalean, kalitate oneko erreferentziazko testuekin konparatuz. Ondoren corpus guztiko emaitzen batezbestekoa lortzen da, kalitate orokorra neurtzeko. Ulergarritasuna edo zuzentasun gramatikala ez dira kontuan hartzen. Hala ere, corpus mailarako pentsatua dago eta ez du emaitza onik ematen banakako esaldiak ebaluatuz gero.

Irteera 0 eta 1 arteko zenbaki bat izaten da. Balio honek itzulpen automatikoa eta erreferentziazkoaren arteko antzekotasuna adierazten du, non 1-etik gero eta hurbilago-

ko zenbakiek antzekotasun handiagoak adierazten dituzten. Eskuzko itzulpenean egon daitezkeen aukera anitzengatik, ia ezinezkoa litzateke BLEU zenbakia 1-era iristea. Horregatik, helburua ez da hori. Hala ere gero eta erreferentzia testu gehiago lortu, zenbakia orduan eta altuagoa izango da.

4. KAPITULUA

Proiektuko proposamena

Aurretik 1.2 atalean aipatu bezala, azken hamarkadetan itzulpen automatikora zuzendutako ikerketak aurrea pauso handiak eman ditu. Pauso horiek estatistikan oinarritutako itzulpenean (SBMT) ageri dira gehien bat, baina gaur egun muga batera iritsi ote den pentsatzen da, hemendik aurrera lortu daitezkeen hobekuntzak zalantzan jartzen hasiz. Honi aurre egiteko azken urteetan erregeletan oinarritutako sistemekin ere lan egin da, bi metodo hauen hibridazioak garatuz.

Itzulpen sistema batean zein bestean ere, hasieratik aurkitzen dira erronka berdinak, gizakiak erraz ulertu baina makinarentzat gaindiezinak direla ematen dutenak, pixkanaka hobetu arren:

- Itzulpena esaldiz esaldiz egiten da, testuingurua ez da kontuan hartzen eta hizkuntza bakoitzak dituen esamolde propioak ere ez. Herrialde ezberdinetan hitz egiten diren hizkuntza beraren aldaerak ere ez dira kontuan hartzen normalean.
- Testuinguruaren arabera hitz bat edo bestea aukeratu behar izanez gero, ez dago ziurtasunik ona aukeratuko dela esateko. Hitz batek gainera, esanahi edo adiera bat baino gehiago izan ditzake.
- Izen bereziak identifikatzea oraindik zaila da, batzuetan gainera entitate izenak oso luzeak izan daitezke, aditzekin, esaldi arrunt bat balitz bezala. Lortzen den itzulpena ez da gizaki batena bezain naturala.
- Hitz bakoitza ongi itzultzea lortuz gero ere, esaldiaren egitura guztia itzuli behar da: ordena ez da berdina, hizkuntza batean hitz bat dena bestean multzoa izan daiteke, aurrizki eta atzizki sistema ezberdinak etab. eta euskararen egitura, ingelesa,

gaztelera edo frantsesarekin konparatuz oso ezberdina da.

Arazo hauek kontuan hartuz, SBMT-n aurkitutako sabaiari aurre egiteko, erregeletan oinarritutako TectoMT garatu da. Honek, esan bezala, sakoneko sintaxia erabiltzen du itzulpena egiteko: esalditik ahal bezainbat ezaugarri atera, hizkuntzarekiko independenteak, eta horiek erabili helburuko hizkuntzako esaldia sortzeko. Honela, esaldiko informazioa duen egiturak hizkuntzarekiko lotura handirik ez duenez, batetik bestera pasatzean parte handi bat mantentzen da, aurkitu diren hainbat arazori erantzuna emanez.

Hain zuzen ere, maila semantikoan definitzen diren rol semantikoek, egitura sintaktiko batetik bestera berdinak diren bezala, hizkuntza batetik bestera ere berdinak izaten jarraitzen dute, gehiengoek behintzat. Gerta daiteke hizkuntza baterako rol batzuk besterako ez existitzea. BVI lexikoian, euskarazko aditzen adiera batzuetarako kasu markak definituta daudela ikusi da. Argumentu bakoitzari rol bat dagokio, eta kasu marka bat edo gehiago (ine, ins, erg etab., ikus 3.3 adibidea). *Mate Tools* edo *Clearnlp* tresnak erabiliz, ingelesezko testuetako hitz bakoitzari dagokion argumentu etiketa jarri dakioke.

Donostiako Euskal Herriko Unibertsitateko IXA ikertzaile taldeko kide batzuek TectoMT itzultzailea hobetzen dihardute, ingelesa-gaztelera eta ingelesa-euskara bikoteetarako. Hau guztia ikusita, proiektu honetarako honako proposamena egin da: analisi semantikoa itzultzaile automatikoan integratzea eta rol semantikoez baliatuz ingelesetik euskarako itzulpena hobetzen saiatzea, TectoMT sistemarako.

TectoMT-n SRL tresna bat integratuz gero, ingelesezko testuan etiketak gehitu daitezke. BVI lexikoiak euskara eta ingelesezko aditz adieren kasu markei buruzko informazioa ematen duenez, ingelesetik euskarara itzultzean baliokidetzak aurkitu, eta euskarazko hitzen kasu markak egokitu litezke. Honela, automatikoki gaizki jartzen diren kasu batzuk zuzendu egingo lirateke.

Gainera, BVI-n aditzen adiera ezberdinak erabil daitezke. TectoMT-k ez du ezberdintasuna ulertzen, eta adiera txarrerako itzulpena proposatzen du batzuetan, lema txarra aukeratuz. SRL tresnek ingelesez aditz adierak etiketatzen dituztenez, BVI-n baliokidetzak bilatuz aditzen itzulpena ere hobetu ahal izango da agian.

TectoMT sisteman aldaketa hauek integratzean datza proiektua. Proiektu esperimentala da; probak egin beharko dira, aurreikusitako hobekuntza horiek zein heinetaraino gertatzen diren ala ez ikusteko. Horretarako, corpus ezberdinekin itzulpenak lortu eta emaitzak bi modutara aztertzea da helburua:

- Azterketa kuantitatiboa: Itzulpenak, eskuz itzulitako erreferentzia batekin konpa-

ratuta automatikoki ebaluatzen dira, aurrez azaldutako BLEU zenbakia ateraz. Rol semantikoak kontuan hartu aurretiko eta ondorengo emaitzak konparatu egingo dira.

- Azterketa kualitatiboa: Eskuz, esaldiz esaldi aztertuko dira aldaketak. Errepikatzen den akats edo emaitza bereziren bat begiz ikusiz gero esaldi horiek aztertuko dira, bestela ausaz edo aldez aurretik ezarritako beste irizpide bat jarraituz.

Emaitza hauek aztertzean, agian zentzurik gabeko aldaketak ikusiko dira. Kodean egindako akatsen bat izan daiteke arrazoia, edo aurreikusita ez dagoen ezusteko bat. Agian akatsik egon ez arren, hobekuntza berri bat pentsatuko da. Orduan, kodea berriz aldatu eta emaitza berriak aztertuko dira. Iterazio hauekin jarraituko da beharrezkoa ikusten den bitartean eta proiektuaren epeak ahalbidetzen duen heinera arte.

5. KAPITULUA

Proiektuaren garapena

Lan hau hiru zati nagusitan banatu da: i) gaian sartzea, itzultzailean eskua sartu eta kodea aldatzen hasi aurretik egin beharrekoa alegia: ezagutza teknikoak eskuratzea, eta SRL tresnak aztertu, saiatu eta aukeratzea; ii) aldaketak finkatu eta erabakitakoa kodetzea eta iii) esperimentazioa eta egindako aldaketekin emaitzak aztertzea, ondorioak ateratzeko. Hala ere, lehentxeago aipatu bezala, emaitzak aztertu ondoren kodea egokitu eta behar bezainbat iterazio egin dira. Atal honetan, garapeneko hiru etapa hauetan emandako pausoak azalduko dira.

5.1 Gaian sartzea

Lehen lana proiektuarekin lotutako gaiei buruz ikastea izan da; hizkuntzalaritzako zein itzulpen automatikoko ezagutza teknikoak eskuratu behar izan dira. TectoMT itzultzailearen egitura eta erabilera ezagutzeaz gain, hau Perl-ez kodetua dagoenez, kodetzeko lengoaia hau erabiltzen ere ikasi behar izan da.

Proiektuarekin hastean, zehazki zein den definitu da: rol semantikoak baliatuz TectoMT-ren itzulpena hobetzea. Prozedura hori nola eraman aztertu behar izan da, pausoz pauso egin beharreko aldaketak, hainbat kontzeptu barneratzearekin batera: i) rol semantikoak zer diren jakin, ii) rolen etiketatzea ezagutu eta iii) BVI, *Mate Tools* eta *ClearNLP* tresnak ezagutu.

Tresna hauen artean *Mate Tools* eta *ClearNLP*-rekin lanketa berezia eraman behar

izan da. Bietako bakarra erabili da, baina hasieran tresnak proiektuko interesatu guztientzat ezezagunak zirenez, bakoitza aztertu behar izan da, bere alde on eta txarrak konparatu, eta irizpide hauen arabera erabakia hartu: erabilerraztasuna, emaitzak lortzeko denbora, lan honetarako erabilgarritasuna eta TectoMT-n integratzeko zailtasun aurreikuspena. Txandaka, probak egin dira bata eta bestearekin, konparaketa osatuz.

5.1.1 *Mate Tools*

Mate tresnak lantzen hasteko, esan beharra dago honi buruzko informazioa urria dela. Dena den, *parser* eta eredu berrienak deskargatzeko atal bat aurkitu da, proiektuaren deskribapenarekin batera. Bertan tresna ezberdinak deskargatu daitezke, lematizatzaile, etiketatzaile morfologiko etab., baita hainbat hizkuntzarako eredu entrenatuak ere. Tresnen artean gida labur bat ere aurki daiteke eta *CoNLL 2009*ko formatuko adibide esaldiak, probak egiteko.

Proiektuko helburua jomugan edukiz, *srl-4.31.tgz* paketea erabili da, ingeleseko ereduekin. *scripts* izeneko direktorioan, erabilgarriak izan daitezkeen bi *script* daude, besteak beste:

- *parse_full.sh*: Honek tokenizazioa izan ezik *Mate*k dituen beste tresna guztiak aplikatzen ditu, lematizatzailea, PoS etiketatzailea, dependentzia *parser*ra eta rol semantikoen etiketatzailea. Bertan dauden azalpenek sarrerako fitxategiak izan behar duen egitura eta *script*a nola egokitu behar den ere argitzen dute.
- *parse_srl_only.sh*: Hemen sarrerako fitxategiak dependentzia formatuan egon behar du, zuzenean rol semantikoen etiketatzea soilik aplikatzeko; hemen ere azalpenak ematen dira *script*a exekutatu ahal izateko. Hau probatzeko *CoNLL 2009*ko formatuan dauden adibideak deskargatu dira, jadanik etiketak dituztenak. 3.3.1 atalean deskribatu diren zutabeetatik azken 4 zutabeetako eremuak azpimarraz ordezkatu eta sarrera fitxategi moduan emanez, irteerak deskargatutako adibidearen berdina izan behar du.

Laguntza honekin, *parse*atzeko modu batean eta bestean emaitza egokiak lortu dira. TectoMT-n dependentziak jadanik jorratzen direnez, denbora aurrezteko egokiena dependentzia formatutik abiatu eta rolak bakarrik etiketatzea da. TectoMT-k ematen dituen datuak *CoNLL 2009*ko datu formatuan jarri eta *parse*atu ahal izango dela aurreikusi da.

5.1.2 *ClearNLP*

Tresna honekin ere, prozedura berdina eramateko saiakera egin da: lehenik, dokumentazioa irakurri erabilera aukera ezberdinak aztertzeko, eta hortik aurrera probak egin. Ez da *Mate* tresnetarako baino askoz informazio gehiago aurkitu. Hala ere, web orrian dagoen informazioa nahikoa da tresnak erabili ahal izateko.

Azalpenak jarraitu ondoren, komando baten bidez, sarrera gisa testu gordina hartuz, *srl* formatuko irteera lortu da. Webguneko beste atal batean, *data format*, formatu ezberdinen adibideak aurki daitezke:

- Formatu gordina: testu guztia batera, ez da formatu berezirik behar.
- Lerro formatua: lerro bakoitzeko esaldi bat.
- Token formatua: lerro bakoitzeko token bat.
- Part-of-speech formatua: Bi zutabeko egitura, lerro bakoitzeko hitzaren forma eta PoS etiketa.
- Morfologia formatua: Hiru zutabe, forma, lema eta PoS.
- Dependentzia formatua: 7 zutabe ; ID, forma, lema, PoS, ezaugarriak, burua eta dependentzia etiketak.
- Rol semantiko formatua: Aurreko 7 zutabeak, argumentu etiketentzat 8. zutabe batekin.

Mate Tools-ekin bezala, hemen ere rol semantikoak etiketatzeko tresna soilik erabil daitekeela pentsatu da. Exekutatu beharreko aginduan helburua zein den zehaztu egiten da, dep, morph edo srl bezalako argumentuekin. Hortaz, abiapuntua ere zehazteko aukera ematen duen beste parametro bat egon daitekeela pentsatzea ez da zentzugabekeria. Gainera, *Penn Treebank* formatuan dagoen dependentzia zuhaitzetik *ClearNLP*-ren zutabekako dependentzia formatura pasatzeko tresna badago. Honenbestez, dependentzia formatu horretatik abiatu eta rol semantikoak etiketa daitezkeela pentsatu da.

Alta, aukera hau ez da inon aurkitu. Ezer zehaztu gabe dependentzia formatuko testu bat prozesatzeko saiakera egin da hala ere eta etiketatzea ondo burutzen duela ikusi da. 3.3.2 atalean deskribatutako zutabeen araberako sarrera fitxategia sortu da, bertako irteera berdina lortuz (3.6 taula). Hona sarrera formatua:

1	The	the	DT	_	4	NMOD
2	most	most	RBS	_	3	AMOD
3	troublesome	troublesome	JJ	_	4	NMOD
4	report	report	NN	_	5	SBJ
5	may	may	MD	_	0	ROOT
6	be	be	VB	_	5	VC
7	the	the	DT	_	11	NMOD
8	August	august	NNP	_	11	NMOD
9	merchandise	merchandise	NN	_	10	NMOD
10	trade	trade	NN	_	11	NMOD
11	deficit	deficit	NN	_	6	PRD
12	due	due	JJ	_	11	APPO

5.1 Taula: ClearNLP-ko dependentzia formatu adibidea

Konparaketa

Rolak etiketatzeko *Mate Tools* eta *ClearNLP* bi tresnak probatu dira, baita biekin lortu ere. Alta, ezberdintasun nabarmenak ikusi dira aukeraketa baldintzatu dutenak.

Alde batetik, *ClearNLP*-ri buruzko dokumentazioa osoagoa eta txukunagoa da, nola erabili ulertzeko errazagoa beraz. Hasieran, *Mate Tools*-ekin saiatzean eta dokumentazioa eskasa dela ikustean, zuzenean *ClearNLP* erabiltzea pentsatu da. Berehala ikusi da ClearNLP-rena ere ez dela hain argia.

Edonola ere, bi tresnak ibilaraztea lortu da. Bigarren ezberdintasuna rolak etiketatzeko abiapuntua da. *Mate*k bi aukera ematen ditu, testu gordin ala dependentzia formatutik abiatzea. *ClearNLP*-k berriz ez du horrelako aukerarik, edo dokumentatuta ez behintzat, baina dependentzia formatuko testu bat prozesatu ezkero, ongi etiketatzeko ditu rolak. Esan bezala, TectoMT-k jadanik dependentziekin lan egiten baitu, eta *CoNLL 2009* formatuarekin hain zuzen ere, puntu interesgarria da lan guztia berregin gabe rolak etiketatu ahal izatea.

Egokia izango zatekeen tresna hauen beste ebaluazio bat egitea, emaitzen kalitatearen konparaketa argiagoa edukitzeko. Halako emaitzarik ez da aurkitu ordea. Hala ere, helburu nagusia TectoMT sisteman integratzeko tresna eroso eta egokiena aukeratzea zen. Hortaz, *Mate Tools* tresna aukeratzea erabaki da.

5.2 Inplementazioa

Atal honetan proiektuaren helburua aurrera eramateko kodean egindako aldaketak azaltzen dira. Proiektuan, rol semantikoak automatikoki etiketatzeko tresna bat erabili nahi izan da TectoMT itzultzailean hobekuntzarik ekartzen duen edo ez ikusteko. Lan hau ingelesetik euskararako itzulpenerako egin da. Honela banatu dira atazak:

Inplementazio azpiataleko egitura

5.2.1 Treex-etik kanpo: BVI moldatzea

5.2.2 Treex-eko aldaketak

5.2.2.1 Ingelesezko testuan rolak etiketatu

- *Mate* tresna, Treex-en gehitu
- Blokeetatik tresnari deia egiteko lotura
- Lortutako etiketak azaleko sintaxiako egituran gorde

5.2.2.2 Ingelesa-euskara informazioa lortu

- BVI-ko informazioa eskuratu
- Euskara-ingelesa nodo baliokideak eta SRL datuak lortu

5.2.2.3 Euskarazko testuan BVI-ko informazioa erabili

- BVI-ko adierak erabiliz, itzulitako hitzen lema aldatu
- Ingelesa-euskara argumentu baliokidetzak
- Argumentu horren BVI-ko kasu marka
- Formeme atributuaren aldaketa

5.2.1 Treex-etik kanpo: BVI moldatzea

Jatorrian, BVI lexikoia aurkezten den moduan, euskarazko aditz adiera bakoitzari dagozkion ingelesezko ordainak azaltzen dira, argumentu bakoitzari dagokion rolarekin. Kasu honetan aldiz, ingelesezko aditz adiera jakin bat emanda honi dagozkion euskarazko aditz eta rolak eskuratu nahi dira. Horretarako, Treex-etik kanpo *script* batzuk prestatu behar izan dira. Hauekin lortutako fitxategi moldatua, Treex-en gorde da, itzulpenean atzigarri egoteko.

Oroitu BVI-ko informazioaren egituraz, ondoko 5.2 taulan ikusgai:

ABESTU						
	1					
	sing_01					
		Arg0	agent	esperimentatzailea	erg	+giz
		Arg1	topic	gaia	abs	-biz/+konkr
		Arg2	recipient	dat		
AGERTU						
	1					
	appear_01/emerge_01					
		Arg1	theme	gaia	abs	
					par	
		Arg2	location	kokapena	ine	

5.2 Taula: BVI-ko informazioaren egitura

Bukaeran lortu nahi den egitura berriz honakoa da:

sing.01			
	abestu		
		A0	erg
		A1	abs
		A2	dat

5.3 Taula: BVI-ko informazioa, moldatu ondoren

Jatorrizko fitxategia lerroz lerro tratatzen da. Ingelesezko adiera jakin baten ordez euskaraz zein aditz erabili daitekeen jakin nahi da. Beraz, aditz horren adierak ez du garrantzirik izango, lemak baizik.

Jarraitzeko, ingelesezko aditza eta haren adiera, biak gorde behar dira. *Mate* tresnak rolak etiketatzean, predikatua « aditza.00 » formatuan etiketatzen du. Hortaz, ondoren tratamendua errazagoa izan dadin, azpimarra puntu batez ordezkatu da. Horrez gain, batzuetan ingelesezko ordain bat baino gehiago egon daitezke argumentu berekin, 5.2 taulan ageri den moduan. Biak tratatu behar dira.

Ingelesezko adieraren ondorengo lerroak 5 zutabetan egituratzen dira: argumentu zenbakia, *VerbNet*eko rola, EADB Euskal Aditzen Datu Baseko rola, deklinabide kasua eta hautapen murriztapenak, ondoko 5.2 taulan ikus daitekeen moduan.

Helburua, aditz adierekin aditzaren lema hobetzea, eta rol semantikoei esker hitzen deklinabidea hobetzea da. *Mate*ek rolak *A0* moduan etiketatzen dituenez, argumentu zenbakitua gorde behar da. Beste zutabeetan, deklinabide kasua da interesgarria den bakarra,

*VerbNet*ek *Agent* edo *Patient* dela esateak ez baitu argumentu zenbakiari dagokion kasua aldatzen. Horregatik, bi zutabe horiek bakarrik mantentzen dira helburuko egiturarako, berriro ere tratamendua errazteko Arg0, A0 bilakatuz. Argumentu batzuek kasu bat baino gehiago eduki ditzakete. Kasu horiek kontuan hartu behar izan dira, denak gordetzeko.

Bestalde, ingelesa eta euskararen arteko trukaketa egin denez, multzoak ere ezberdinak izan daitezke, hau da: lehen *AGERTU* aditzaren adiera baten barne *appear.01* zegoen, baina baita *AZALDU* aditzaren barne ere. Orain, *appear.01* aditzaren barne *agertu* eta *azaldu* egongo dira, bakoitza bere argumentuekin.

Laburbiltzeko, adibide batekin argiago ikusten dira aldaketa guztiak. Lehenik, 5.4 eta 5.5 tauletan jatorrizko informazioaren parte bat ageri da, euskarako aditzak eta hauen ordainak ageri direlarik:

AGERTU					
	1				
		appear_01/emerge_01			
			Arg1	theme	gaia	abs
			Arg2	location	kokapena	par ine
	2					
		appear_02				
			Arg1	theme	gaia	abs
			Arg2	predicate	egoera	abs soz mod
	3					
		show_01				
			Arg0	agent	esperimentatzailea	erg
			Arg1	topic	gaia	abs
			Arg2	recipient		dat

5.4 Taula: BVI-ko jatorrizko egitura

AZALDU					
	1				
		explain_01/state_01/demonstrate_01			
			Arg0	agent	esperimentatzailea	erg
			Arg1	topic	gaia	abs
						konpl
			Arg2	recipient		dat
	2					
		appear_01				
			Arg1	theme	gaia	abs
			Arg2	location	kokapena	ine
						ala
	3					
		appear_02				
			Arg1	theme	gaia	abs
			Arg2	predicate	egoera	abs
						soz

5.5 Taula: BVI-ko jatorrizko egitura

Orain, 5.6 taulan amaieran lortzen den egitura ikus daiteke. Hemen, ez dago informazio guztia, ez baitaude *emerge_01*, *appear_02* eta beste adierak ere:

appear.01			
	agertu		
		A1	abs
			par
		AM-LOC	ine
	azaldu		
		A1	abs
		AM-LOC	ine
			all

5.6 Taula: BVI-ko amaierako egitura

Esperimentazio bitartean datu hauetan akats eta ezustekoak aurkitu dira, hasieratik ikusten ez zirenak. Aldi oro fitxategia bera zuzendu eta moldatu da, Treex-en zuzenean tratagarria izan dadin, aldaketarik egin gabe. Treex-en gordetakoa eguneratu da aldatutakoa utziz:

- Euskarazko kasu batzuk TectoMT-k etiketatutako ezberdinak ziren, beraz baliokidetzak kontuan hartu dira. TectoMT-k etiketatutakoa aldatu ezin denez, euskarazko kasuak TectoMT-koak bezala jarri dira. Honela, 'ala' kasua 'abl' bilakatzen da, 'soz' 'com' bihurtzen da, eta beste hainbat ere aldatzen dira.

- Ingelesez etiketatutako argumentu zenbaki eta BVI-ko argumentu zenbakietan ere arazoak aurkitu dira, ez baitagozkie zenbaki berdinak rolei.

Ondoko adibidean errazago ulertuko da bigarren akatsa:

(8) *Try moving the pointer to a different area and see if it reappears.*

Kasu honetan, punteroa da mugitzen dena, zehaztuta ez dagoen leku batetik eremu ezberdin batera. Beraz, kasu ablatiboa (NONDIK adierazten duena) ez da agertzen eta adlatiboa (NORA adierazten duena) bai, eremu ezberdina litzateke.

Ondoko 5.7 taulan ikus daiteke hasieran BVI-tik ateratako informazioetatik Treex-erako finkatutako formatua. Aipatutako kasuen abrebiazioan, bi hauek hurrenez hurren 'abl' eta 'all' dira eta hauen argumentu zenbakiak A2 eta A3.

move.01		
	mugitu	
	A0	erg
	A1	abs
		par
	A2	abl
	A3	all

5.7 Taula: *move.01* aditzaren rolak, azken moldaketaren aurretik

Honen ondorioz, (8) adibideko *different area*-ren itzulpenak, adlatiboak alegia (nora), A3 argumentua izan behar luke. Treex-en itzulpena egitean ordea, *area* A2 argumentu bezala etiketatu du, gure kasuan ablatiboari dagokiona (nondik). *Mate*k etiketak beti ongi jartzen ez dituela suposatu arren, ingelesezko *move to*+... forma agertzean etiketa ablatiboa jartzea akats larria litzateke, oinarrizkoa. Horregatik, tresnaz kanpo beste ezohikoren bat zegoela ondorioztatu da.

Honela, informazio iturrietan pixka bat bilatuz, euskaraz eta ingelesez argumentu zenbakiak ezberdinak direnaz ohar daiteke. Ondoko 5.8 taulan *mugitu* hitzaren ingelesezko ordainak BVI-n eta *PropBank*-en dituen argumentuak ageri dira:

move.01		*PropBank*, move.01:
mugitu		Arg0-PAG: mover
A0	erg	Arg1-PPT: moved
A1	abs	Arg2-GOL: destination
	par	
A2	abl (nondik)	
A3	all (nora)	

5.8 Taula: *move.01* aditzaren rolak euskaraz eta ingelesez *PropBank*-en arabera

Ingelesez, *PropBank*-en arabera, jatorria (nondik) ez da kontuan hartzen argumentu zenbakietan, beraz, euskaraz BVI-n jatorria adierazten zuen argumentuak hemen helburua adierazten du. *Mate*k beraz, ingelesezko aditzak *to...* forma duenean etiketa egokiak jartzen dizkio, baina gure BVI-ko informazioak ez ditu baliokidetza onak.

Arazo hau konpontzeko, IXA taldeko hizkuntzalari batzuen laguntza baliatu da. Hauek, arazo horiez jadanik konturatuta, baliokidetza batzuk prestatu eta helarazi dituzte, ondoko itxuran:

moztu.1	0	block.01	0
moztu.1	0	cut.01	0
moztu.1	0	cut.01	1
moztu.1	1	block.01	1
moztu.1	1	cut.01	1
moztu.1	2	block.01	1
moztu.1	2	cut.01	1
moztu.1	3	cut.01	3
mugitu.1	0	move.01	0
mugitu.1	1	move.01	1
mugitu.1	**3**	**move.01**	**2**
nahasi.1	0	jumble.01	0
nahasi.1	1	jumble.01	1
nahasi.1	2	jumble.01	1

5.9 Taula: Euskarazko argumentu zenbakiek dagozkien ingelesezkoen adibide batzuk

5.9 taulan ingelesa eta euskara arteko argumentu zenbakien baliokidetza adibide batzuk ageri dira. Lehen zutabeko euskarazko aditz adierari dagokion bigarren zutabeko argumentu zenbakiari, 4. zutabeko argumentu zenbakia dagokio, 3. zutabeko aditz adierara itzultzean. Hau da, *moztu* aditza *block*-era itzultzean, 0. argumentua berdina izango dute. *cut*-era itzultzean aldiz, *cut*-en 0 eta 1 argumentuetan kasu berdinak izango dira, *moztu*-ren 0. argumentuan daudenak.

Treex-en gordetzen den fitxategia moldatzean ingelesezko argumentuen arabera jartzea erabaki da, lehen aipatu bezala, *Mate*k egiten duenaren arabera gelditzeko eta exekuzioan bertan moldaketarik egin behar ez izateko.

Uneko adibidean, euskaraz *mugitu*-ren 3. argumentuko kasua *move.01*-eko 2. argumentuari dagokio. Hau da, kasu adlatiboa (nora) *destination* informazioarekin lotzen da. BVI ondoko 5.10 taulan ageri den moduan aldatzen da:

Jatorrizkoa		Baliokidetzak	Helburukoa	
move.01			move.01	
mugitu			mugitu	
A0	erg	$\longrightarrow 0 \longrightarrow 0 \longrightarrow$	A0	erg
A1	abs	$\longrightarrow 1 \longrightarrow 1 \longrightarrow$	A1	abs
	par			par
A2	abl	**$3 \longrightarrow 2 \longrightarrow$**	**A2**	**all**
A3	**all**	$\nearrow$ $2? \rightarrow 3? \rightarrow$	$\emptyset$	

5.10 Taula: move.01 aditzaren rolak, aldaketaren aurretik eta ondotik

5.2.2 Treex-eko aldaketak

Aurreko atalean Treex-etik kanpo egindako aladaketak aipatu dira, hau da, BVI-ko informazioaren moldaketa. Zati hau berriz azpimultzo gehiagotan banatzen da: i) ingelesezko testuan *Mate* tresnari esker rolak etiketatuko dira, ii) ingelesa eta euskarazko hitzen arteko loturak egingo dira eta iii) euskarazko testuan BVI-ko informazioa erabiliko da kasu marka edo lemak aldatu eta hobetzen saiatzeko.

5.2.2.1 Ingelesezko testuan rolak etiketatu

Mate tresna, Treex-en gehitu

3.4.1.4 atalean azaldu da Treex-en kanpoko tresnak erabiltzeko aukera dagoela, Tools atalean hauek integratuz, orokorrean Perl ez den besten lengoaia bat erabiltzen baitute. Hau gertatzen da *Mate Tools*-ekin, *Java*n inplementatuta dago eta.

Lehen gauza beharrezkoak diren liburutegi, eredu eta *jar* fitxategiak Treex-era kopiatzea izan da, horretarako sortu den *share/installed-tools/mate-tools* direktoriora. Beraz, bertan, Mate-i buruzko azalpenetan aipatu diren tresnak daude:

- CoNLL2009-ST-English-ALL.anna-3.3.srl-4.1.srl.model

- lib: anna-3.3.jar, liblinear-1.51-with-deps.jar
- srl.jar

Ondotik, bi modulu kodetu dira: 1. *Java*rekin lotura egin eta tresnari deitzeko komandoa sortzeko, eta 2. tratatzeko testu bat jasotzean, lehenengo moduluari dei egin eta tresna erabili daitekeela frogatzeko. Ondoko puntuetan azaltzen dira bi moduluak:

1. *ParseSRL.pm*: Modulu honetan, *Mate* erabiltzeko ereduak, liburutegi eta JAR fitxategiak nondik eskuratu definitzen da. Aipatu moduan *share/installed-tools/matetools*-en daude. Hauek beharrezko beste hitzekin kateatuz, rol etiketatzailea martxan jartzeko *java* deia osatzen da. *parse_document* izeneko funtzioan, esaldi zerrenda bat jaso eta bakoitzeko aginduari dei egiten zaio. Emaitzak ere esaldiz esaldi jasotzen dira. Jasotako esaldiek, esan bezala, *CoNLL*-ko formatuan egon behar dute, irteerakoak ere hala egon daitezen. Funtzio honek irteerako esaldi zerrenda itzultzen du.

2. *t/matetools_parseSRL.pm*
   Fitxategi honetan, probak egin dira. Esaldi zerrenda bat sortu da eskuz, CoNLL 2009ko probarako esaldietatik. Honekin, *ParseSRL.pm*-ri dei egin zaio agindua hasieratzeko, jarraian *parse_document* funtzioari deituz, esaldi zerrenda konkretua argumentu gisara eman eta analizatu dezan. Emaitzak behar dutenak direla egiaztatu arte erabili da modulu hau.

Blokeetatik tresnari deia egiteko lotura

Behin *Java*rekin lotura egin denean, tresna Treex-en integratu eta erabiltzeko aldaketak egin dira. Aipatu moduan sistemak dependentzia formatua lantzen du, baita *CoNLL*-ko ezaugarriak ere. *Block/A2A/EN/MateSRL.pm* blokean *Mate* eta Treex-en arteko dependentzia formatuen egokitzapena egiten da.

Oroitu Treex-en egituraz: *A2A/EN* direktorioak, geruza analitikotik mugitu gabe, hizkuntza bateko testuaren gainean aldaketak aplikatzen ditu, kasu honetan ingelesez dagoena. Bloke honetatik beraz, jasotako esaldi zerrenda erabiliz *ParseSRL.pm*-ko *parse_document* funtzioari deitzen zaio.

Bertan, esaldi bakoitzari dagokion zuhaitza ere eskura dago. Jasotako emaitzak tokenez token eta zutabeka banantzen dira, informazio bakoitza token bakoitzari dagokion nodoko atributuetan gordetzeko (forma, lema, PoS etiketa… printzipioz aurretik ziren berdinak).

Lortutako etiketak azaleko sintaxiako egituran gorde

Ondotik, esperimentuan ekarriko diren benetako aldaketak hasten dira: TectoMT-n, ingelesezko a-zuhaitzeko nodoek, lema, forma, denbora edo dagozkien beste eremuez gain, *wild_dump* eremua dute. Bertan, eskuz, nahi diren eremuak gehitu daitezke, hash taula modura. Predikatuaren nodoari *srl* eremu bat gehitzea erabaki da inguruko hitzen informazio guztia biltzeko.

(9) *Those complaints don't appear to bother Apple.*

(9) adibideko *appear* hitzari dagokion *srl* eremua honakoa da:

```
'srl'   =>{
        'pred' =>'appear.02',
        'A1' =>'2',
        'AM-NEG' =>'4',
        'C-A1' =>'6'
        }
```

'pred' gakoak aditza eta adiera gordetzen ditu. Beste gakoak argumentu zenbakiak dira, eta dagokien hitzaren esaldiko ordena zenbakia gordetzen dute, nahiz eta batzuetan hitz bat baino gehiago izan. Hemen, 2. hitza A1 argumentua da.

Rolak etiketatzean lortutako emaitzetan, PRED zutabeko eremua betea duen token bakoitzeko, predikatua gordetzen da. Predikatu bezainbat zutabe gehitzen dira amaieran argumentuentzat, aditzek testuan duten azalpen ordenaren arabera. Hortaz, zutabe bakoitzeko informazioa zein predikaturi dagokion jakin daiteke, ikus 5.11 taula.

1	Those	_	those	_	DT	_	_	_	2	_	NMOD	_	_	_	_
2	complaints	_	complaint	_	NNS	_	_	_	3	_	SBJ	_	_	A1	A0
3	do	_	do	_	VBP	_	_	_	0	_	ROOT	_	_	_	_
4	n't	_	not	_	RB	_	_	_	3	_	ADV	_	_	AM-NEG	_
5	appear	_	appear	_	VB	_	_	_	3	_	VC	Y	appear.02	_	_
6	to	_	to	_	TO	_	_	_	5	_	OPRD	_	_	C-A1	_
7	bother	_	bother	_	VB	_	_	_	6	_	IM	Y	bother.01	_	_
8	Apple	_	apple	_	NNP	_	_	_	7	_	OBJ	_	_	_	A1
9	.	_	.	_	.	_	_	_	3	_	P	_	_	_	_

5.11 Taula: Predikatu bakoitzari dagozkion argumentuak zutabeetan

Treex-eko nodoek ordena zenbaki bat dute. *Mate*k tokenei ematen dizkien ID zenbakiak ordenaren araberakoak dira, orduan zenbaki berdinak dituzte tresna batean eta bestean.

Argumentuen zutabeetan, eremu bat betea dagoenean, dagokion ID zenbakia eta argumentua gorde behar dira, predikatuari dagokion ingelesezko a-nodoko *wild_dump* eremuan. Predikatuari dagokion nodoa berriz, tokenaren ID zenbakiaren ordena berdina duen nodoa izango da, 5.12 taulan ageri den moduan.

1	Those	_	those	_	DT	_	_	_	2	_	NMOD	_	_	_	_
2	complaints	_	complaint	_	NNS	_	_	_	3	_	SBJ	_	_	A1	A0
3	do	_	do	_	VBP	_	_	_	0	_	ROOT	_	_	_	_
4	n't	_	not	_	RB	_	_	_	3	_	ADV	_	_	AM-NEG	_
5	appear	_	appear	_	VB	_	_	_	3	_	VC	Y	appear.02	_	_
6	to	_	to	_	TO	_	_	_	5	_	OPRD	_	_	C-A1	_
7	bother	_	bother	_	VB	_	_	_	6	_	IM	Y	bother.01	_	_
8	Apple	_	apple	_	NNP	_	_	_	7	_	OBJ	_	_	_	A1
9	.	_	.	_	.	_	_	_	3	_	P	_	_	_	_

5.12 Taula: Argumentu eta ID zenbakiaren arteko lotura

Adibide honetan *appear* eta *bother* hitzek izango dute *srl* eremua, hurrenez hurren, honelakoa:

```
'srl'   =>{
        'pred' =>'appear.02',
        'A1' =>'2',
        'AM-NEG' =>'4',
        'C-A1' =>'6'
        }
'srl'   =>{
        'pred' =>'bother.01',
        'A0' =>'2',
        'A1' =>'8',
        }
```

5.2.2.2 Ingelesa-euskara informazioa lortu

Hemendik aurrerako lana *Block/T2T/EN2EU/ApplyRolRestrictions.pm* blokean kodetu da. Bloke hau, beste asko bezala, esaldi bakoitzeko exekutatzen da. Aldi bakoitzeko, euskarazko t-nodo bat jasotzen da, beraz, nodoz nodo exekutatzen da.

BVI-ko informazioa eskuratu

Hasteko, BVI-ko informazioa eskuratu behar da, esan bezala proiektuko beharretara egokitu dena. Aipatu bezala, moldaketa bera Treex-kin kanpo egin arren, tresnen moduan emaitza Treex-en gorde behar da. *bvi_en2eu_ordainak.txt* izenpean, *share/data/models/BVI/* direktorioan kokatzen da.

Fitxategi hau lerroz lerro irakurri eta aldagai global batean gordetzen da exekutatutako lehen aldian. Honela, ez dago denbora alferrik galdu beharrik esaldi bakoitzeko. Hash taula batean ingelesezko aditz adiera, haren euskarazko ordaina, argumentu eta kasu marka ezberdinen informazioa gordetzen da.

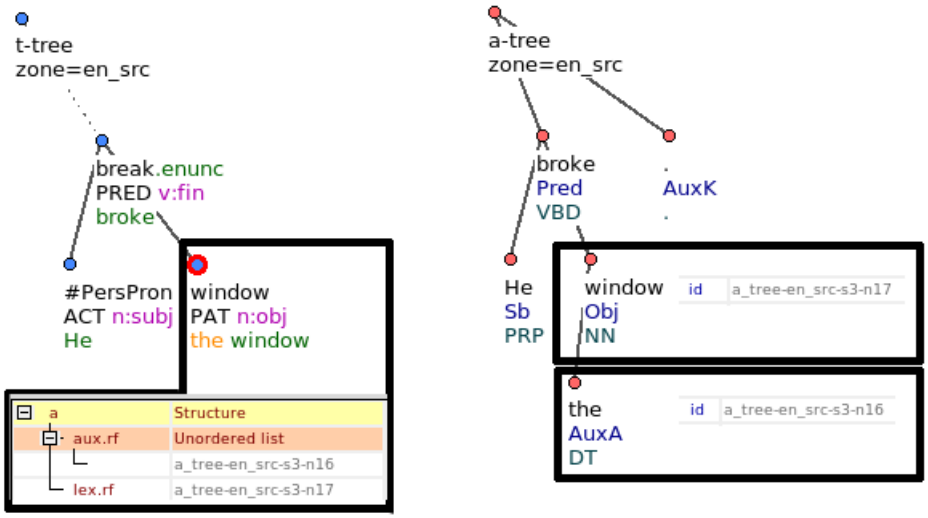
Euskara-ingelesa nodo baliokideak eta SRL datuak lortu
Bloke honek ingelesetik euskarara itzultzeko momentuan, geruza tektogramatikalean aldaketak eragiten ditu. Euskarazko t-geruza lantzen da, eta dagokion jatorrizko ingelesezko t-geruza. Azken honetatik hala ere, ingelesezko a-geruzako informaziora heldu daiteke.

Rol semantikoen inguruan jarritako etiketak ingelesezko a-nodoetan gorde direnez, hauen eta euskarazko t-nodoen arteko lotura lortu behar da. Horretarako, euskarazko t-nodo bati dagokion ingelesezko t-nodoari dagokion a-nodoa lortzeko funtzioak daude:

```
$src_tnode = $tnode->src_tnode();
@anodes = $src_tnode->get_anodes();
```

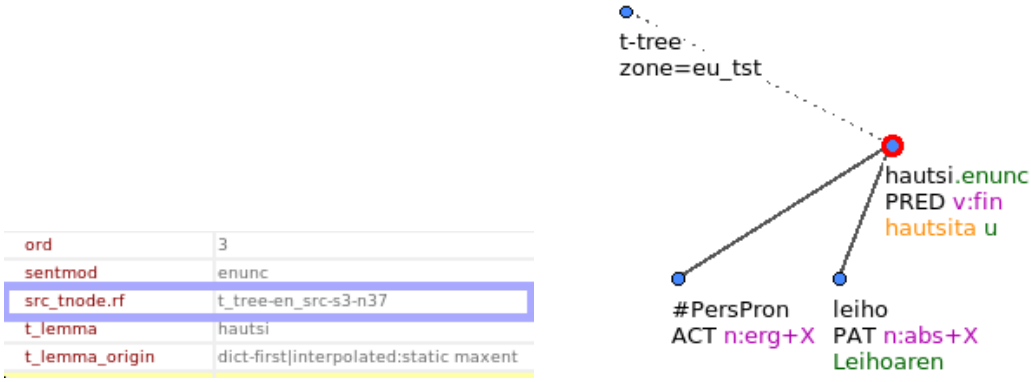
Kode lerro hauetan, $tnode euskarazko t-nodoa da, $src_tnode jatorrizkoa, ingelesezkoa alegia, eta @anodes, $src_tnode-ri dagozkion a-nodoak.

T-nodo baten a-nodo baliokideak, bat baino gehiago izan daitezke. Oroitu, t-nodoak esanahia duten hitzak bakarrik direla, eta hauen laguntzaileak atributu gisara gordetzen direla; a-zuhaitzetan hitz bakoitzak du nodoa. Adibidez, ondoko 5.1 irudian *the window*, t-zuhaitzean nodo bakarra izango da, a-zuhaitzean berriz bi.

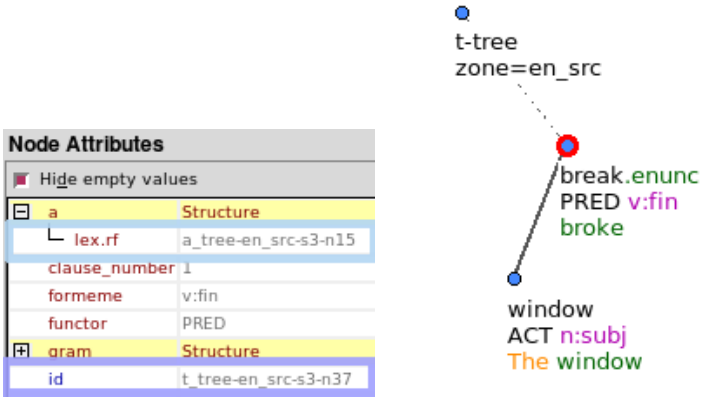


5.1 Irudia: t-nodo bati dagozkion a-nodoak

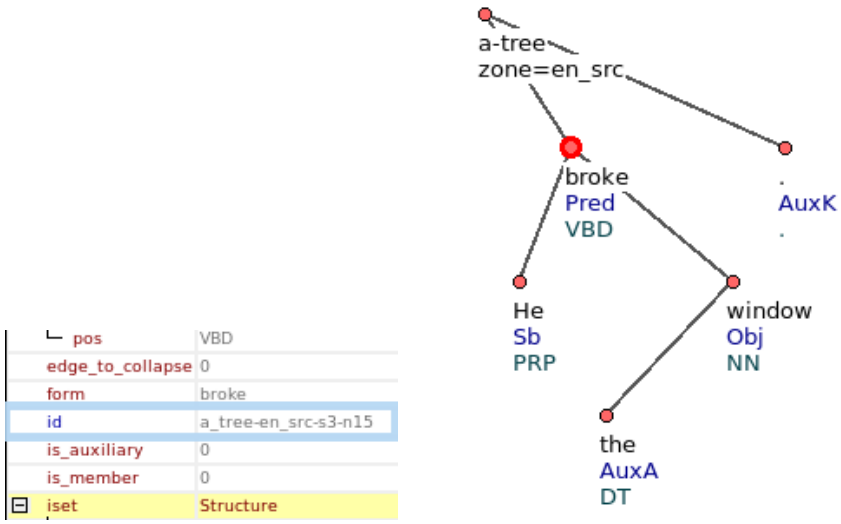
Nodo batetik besterako lotura hobeto ikus daiteke ondoko 5.2, 5.3 eta 5.4 irudien bitartez, adibide honetan aukeratutako nodoa aditza izanik (gorriz inguratutakoa):



5.2 Irudia: Euskarazko t-nodoa eta jatorrizko t-nodoaren arteko lotura



5.3 Irudia: Ingelesezko t-nodoaren eta haren a-nodoen arteko lotura



5.4 Irudia: Jatorrizko a-nodoa eta bere identifikatzailea

Euskarazko eta ingelesezko nodoen arteko lotura egin ondoren, ingelesezkoek duten informazioa eskuratu behar da.

Nodo guztiek ez baitaukate rol semantikoei buruzko informazioa, euskarazko t-nodoari dagokion ingelesezko a-nodoa eskuratu eta *wild_dump* eremuan *srl* azpi-eremua duen edo ez begiratzen da. Hala bada, bertako informazio guztia gordetzen da.

5.2.2.3 Euskarazko testuan BVI-ko informazioa erabili

Puntu honetan, euskarazko t-nodo eta ingelesezko a-nodo bakoitzaren arteko lotura egiteko modua dago, eta *srl* eremua dutenetan informazioa lortuta ere bai. Hemen, t-nodoak aipatzerakoan euskarazkoak izango dira, eta a-nodoak ingelesezkoak.

BVI-ko adierak erabiliz, itzulitako hitzen lema aldatu
Treex-ek egindako nodoaren itzulpena egokia den ala ez jakiteko ez dago modurik. Sinonimoak diren bi hitzen artean bat edo bestea aukeratzeak ez du garrantzia handirik. Aldiz, adierak kontrolatu gabe, esanahia aldatzeko arriskua dago.

Sistemak itzulpena egiterakoan, hitzari lema jartzeko aukera ezberdinak ditu. Hauei, itzulpen egokia izateko probabilitate bat esleitzen die. Lema zerrenda, *t_lemma_variants* deiturikoa, probabilitate hauen arabera ordenatzen da, handienetik txikienera, zerrendako lehena aukeratuz itzulpenerako. Hala ere, zerrenda guztia gordeta gelditzen da.

Hemen, helburua zerrenda hori BVI-ko aditzekin konparatzea izan da. Ingelesezko aditz batek adiera bat baino gehiago eta adiera bakoitzak euskarazko ordain bat baino gehiago izan ditzake. Rolak etiketatzean adiera zehazten denez, ordain posibleak mugatzen dira. Esan bezala, BVI-ko informazioa aldagai global batean gordetzen da hasieran, aldi oro fitxategia irakurri behar ez izateko.

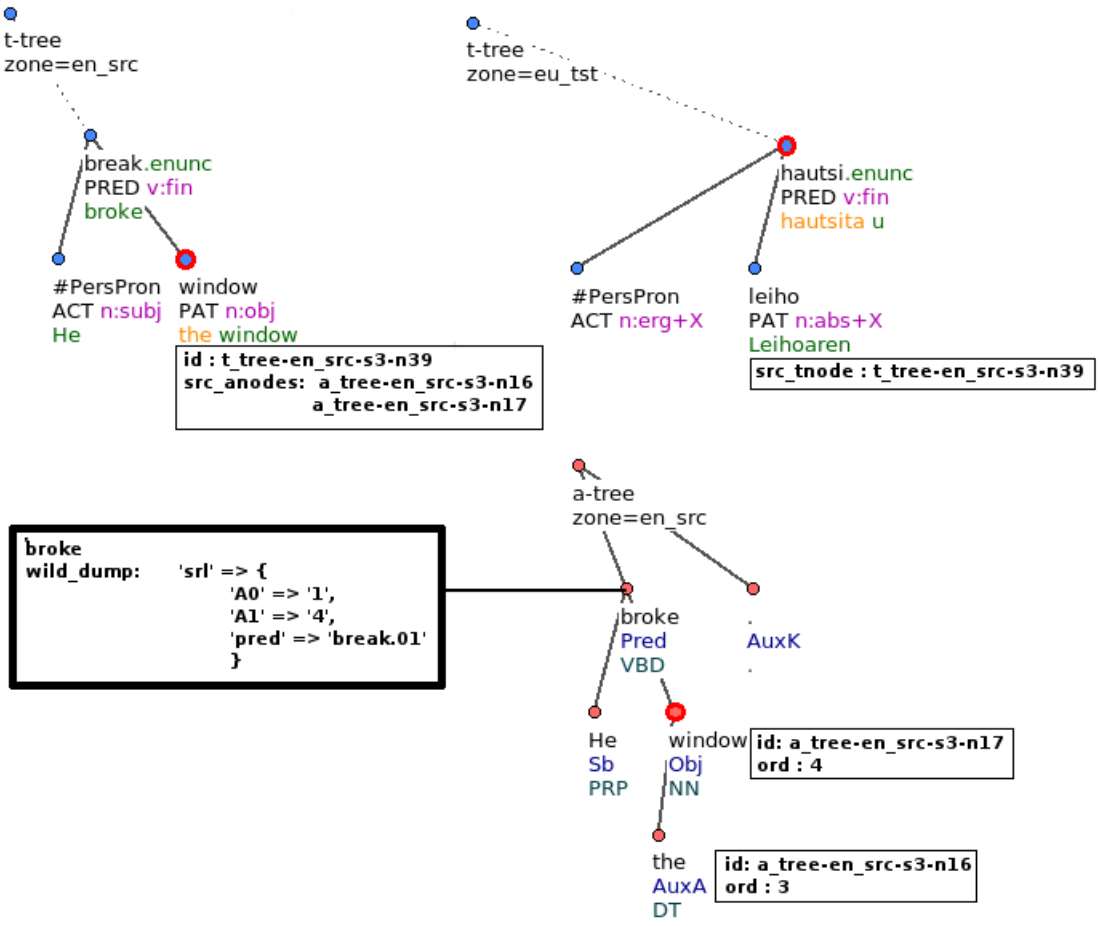
t_lemma_variants zerrendako lemak, dauden ordenan banan banan aztertu dira. Bakoitzeko, rolak etiketatuta lortu den ingelesezko aditz adiera eta uneko lema bikotea BVI-n bilatzen da.

Bikote hori existituz gero, zerrendako uneko lema euskarazko t-nodoaren lemari esleitzen zaio. Lehenengoa bada, ez du aldaketarik ekarriko. Hala ez bada, hurrengo lema proposamenarekin jarraituko da, baten bat aurkitu edo zerrenda amaitu arte.

Ingelesa-euskara argumentu baliokidetzak
Aditzari esker lortutako SRL informazioa, inguruko hitzetan erabili da ondoren. t-zuhaitzetan,

ongi sortuta badaude behintzat, predikatuaren argumentu diren nodoak, aditz nodoaren umeak dira. Beraz, bilaketa hauetara mugatzen da.

Euskarazko aditz nodoa aztertzerakoan, blokearen exekuzio horretan alegia, umeak banan banan korritu eta bakoitzari dagokion ingelesezko a-nodoa lortu behar da. a-nodoaren ordena zenbakia, SRL informazioko argumenturen baten balioan agertuz gero, argumentu hori euskarazko t-nodoari dagokiona izango da. Adibide batekin uler daiteke errazago:



5.5 Irudia: t-nodo bati dagozkion a-nodoak

Irudian, *leiho* hitzaren a-nodoak *the* eta *window* dira, 3 eta 4 ordena dutenak. *srl* eremuan A1 argumentuaren balioa 4 da. Beraz, *leiho* hitza ere A1 argumentua izango da.

Argumentu horren BVI-ko kasu marka

Euskarazko t-nodoei dagokien argumentua eskuratuta, BVI-ren arabera argumentuari da-

gokion kasu marka zein den jakin behar da, sistemak automatikoki emandakoa aldatu ala ez erabakitzeko.

Aditzaren ume t-nodo bakoitza zein argumentu den ezagutzen denez, aldagai globalean ingelesezko aditza, euskarazko ordaina eta argumentu zenbakiarekin, kasu marka aukerak eskuratzen dira, bat edo gehiago izan daitezkeenak. Hirukotea existitu ezean, formema zegoen bezala uzten da, eta bestela aurrera egiten da.

Formeme atributuaren aldaketa

Lemekin gertatzen den bezala, Treex-en euskarazko t-nodo bati formema atributua esleitzeko zerrenda bat dago, *formeme_variants* izenekoa. Bertan, probabilitate handienetik txikienera ordenatzen dira aukerak. Ordena horretan, banan banan, formeman agertzen den kasu marka, eskuratu berri den BVI-ko kasu marka aukeretakoren batekin bat datorren begiratzen da.

Lehena bada, formeman ez du aldaketarik ekarriko, sistemak aukeratzen duen bera baita. Bestela, eguneratu egingo da.

6. KAPITULUA

Esperimentuak eta emaitzak

Atal honetan, kodea moldatu ondoren egin den esperimentazioa azaltzen da. Lehenik, emaitzen azterketa kuantitatiboa egin nahi izan da, BLEU zenbakiak baliatuz. Ondorio argirik lortu ez eta eskuzko azterketa egitea erabaki da gero. Pixkanaka, itzulpen arraroak aurkitu dira. Horien jatorria aurkitu eta konpontzeko beste proba batzuk finkatu dira.

Itzulpenak ingelesetik euskarara egin dira. Horretarako prestatutako eszenario bat erabili eta rol semantikoekin lotura duten blokeak gehitu dira, aurreko 5 kapituluan azaldutako aldaketa guztiak dituztenak.

6.1 Corpusak

Esperimentu guztiak bi corpusekin egin dira. Hasieran, rol semantikoek ekarritako informazioa kontuan hartu gabe corpus bakoitzeko itzulpen bana egin da. Ondoren, egindako aldaketak aztertzeko, esperimentu bakoitzeko, berriro bi corpusak itzuli dira: *batch2a* eta *news*.

Ondoko 6.1 taulan corpus hauen tamainak ikus daitezke eta bakoitzaren azalpen labur bat jarraian:

	Esaldiak	**EN hitzak**	**EU hitzak**
batch2a	1000	21148	16706
news	1104	22030	18030

6.1 Taula: *bathc2a* eta *news* corpusen tamainak

- *batch2a* informatikako domeinuko corpusa da. Bertan, gai horrekin lotutako esaldiak daude, argibideak ematen dituztenak.

  (10) a. *In the Insert menu, select Picture.*
  b. *Click on the part of the document where you want the chart and then in the Insert menu choose Object and click where it says graph.*
  c. *You can download it at this site http://Notepad-plus-plus.org/*

  Hauek itzuli beharreko esaldiak dira, baina badaude euskarazko esaldiak ere, erreferentzia gisara hartzen direnak, 3.4.2 atalean azaldutako BLEU zenbakia kalkulatzeko erabiltzen direnak. Hauei esker itzulpena eta erreferentziaren arteko ezberdintasunak ikus daitezke. Hona (10) adibideko esaldien erreferentziak:

  (11) a. Txertatu menuan, hautatu Irudi bat.
  b. Egin klik grafikoa sartu nahi duzun lekuko dokumentuaren zatian eta Txertatu menuan hautatu Objektua eta egin klik grafikoa dioen tokian.
  c. Deskargatu webgune honetan: http://Notepad-plus-plus.org/

- *news* berriz berriei buruzkoa da, izenak dion bezala. Edozein motako artikuluetako testua dago bertan:

  (12) a. *Hundreds of police officers were involved, some of them wearing riot helmets.*
  b. *And two people chained themselves to trees, Mr. Kelly said.*
  c. *But no 18-year-old should be subject to such intimidation and vitriol in an educational institution.*
  d. *Some lawmakers also want to allow for the nullification of federal laws if they are opposed by two-thirds of the states.*

  Hemen ere euskarazko erreferentzia esaldiak aurkitzen dira, (12) adibiderako ondorengoak:

  (13) a. Ehunka poliziak parte hartu zuten; batzuek istiluetarako kaskoak zeramatzaten.
  b. Kellyren hitzetan, beste bi pertsona zuhaitzetara kateatu ziren.
  c. Baina 18 urteko inork ez luke inoiz halako jazarpen eta gogortasunik jasan behar hezkuntza-erakunde batean.

d. Legegile batzuek lege federalak baliogabetu daitezen ere nahi dute, betiere, estatuen bi heren kasuan kasuko legearen kontra badaude.

6.2 Esperimentazioa

Atal honetan esperimentazioa nola egin den azaltzen da. Hasierako helburua BLEU zenbakien azterketa bidez egitea zen, eta honela hasi da. Pixkanaka ordea azterketa kuantitatibo horretan aldaketa esanguratsurik ikusi ez eta azterketa kualitatibora jo da. Lau probatan banatzen da atal hau, lehena BLEU-aren azterketa hutsa izanik eta ondorengoak pixkanaka aurkitutako ezohikotasunetatik abiatuta.

6.2.1 Lehen probak: BLEU zenbakien azterketa

Kodean egokitu ondoren, *Mate*kin rol semantikoak etiketatu eta informazioa BVI-koarekin alderatuta, hitzen lemak aldatzea lortu behar litzateke, adierak hobetuz. Horrez gain, esaldian aditzen inguruan dauden beste hitzen kasu markak ere hobetu behar lirateke.

batch2a informatikari buruzko corpusak domeinu zehatza duenez, BVI-n aditzen ordainak eta adiera egokiak aurkitzeko zailtasunak izango direla aurreikusi da. Izan ere, gauza jakina da BVI-n aditz guztiak, ezta hauen adiera guztiak ere ez daudela eta informatikako aditz batzuk agian ez direla behar bezala egongo, domeinu orokorrekoekiko ezberdinak baitira.

(14) *Start your computer.*

(14) adibide honetan esaterako, *start* hitzak konputagailua piztu edo abiaraztea adierazten du. Testuinguruari kasurik egin gabe ordea, *hasi* hitzera itzul genezake.

news corpusean berriz informatikari buruzkoan baino esperantza handiagoa jarri da lehenengo emaitzei dagokienez, nahiz eta froga egin aurretik ezin den jakin.

Rol semantikoak kontuan hartu aurreko eta ondorengo emaitzak konparatzeko beraz, bi corpus hauen gainean bina itzulpen egin dira, aurreko azpiatalean esan bezala; 4 itzulpen orotara.

Itzulpena ingelesetik euskarara egin da, jadanik finkatutako eszenario eta corpusak erabiliz. Amaieran BLEU zenbakia lortzen da, beste datu batzuen artean. Txandaka, proiek-

tuarekin hasi aurreko zenbakiekin konparatu dira. *roles=1* edo *roles=0* etiketak adierazten du proiektuko aldaketak, rol semantikoekin lotutakoak alegia, aplikatu diren edo ez.:

	roles = 0	**roles = 1**
batch2a	19.71	19.6
news	2.06	2.05

6.2 Taula: 1. esperimenturako BLEU-aren azterketa

Rol semantikoen azterketa aplikatzean BLEU zenbakia txikitzen dela ikus daiteke, *batch2a* corpusean argiago ageriz. Hala ere, ez dago ezberdintasun handirik. Aipatu bezala, *batch2a* corpusarekin hobekuntzak egotea zalantzan jarri da hasieratik, baina okertzea ez hainbeste. Izan ere, domeinu zehatza du eta BVI ez dago horretarako pentsatuta. Lortutako emaitzak hobeto ulertzeko, esaldi batzuk banan banan aztertu dira.

Bi corpusetan lehen begiratuan kasu marketan hobekuntza bat egon litekeela ematen du, baina ez da hain agerikoa. Lemak aldiz berdintsu edo okerrago gelditzen direla dirudi.

Hala eta guztiz ere esan beharra dago, proiektu honen parte ez den arren, jende gehiago ari dela lanean TectoMT itzultzailean. Esperimentazio honekin hasterako orduan, sorkuntza fasean arazo batzuk zeuden, beraz lortutako BLEU zenbaki eta esaldien egiturak ez dira esanguratsuak. Rol semantikoen erabileraren eragina aztertzeko, TrEd interfazeari esker hitzen ezaugarriak begiratu behar dira. Lehen proba honetan hala ere ez da TrEd erabili, begirada orokor bat eman nahi baitzitzaien emaitzei.

Ondoko adibidean rolen azterketa egin aurretik eta ondoreko itzulpenak ageri dira. Ezberdintasun nabarmenenak hiztegi aldaketan daude, BVI-n aurkitutako adierengatik. Horrez gain ez da kasu marketan ezberdintasun handirik nabaritu, sorkuntza txarra dela eta:

(15) a. Gunean hau erabili **da**: http://goo.gl/ gunean. Itsatsi esteka eta **lortu** beste laburra.
Gunean hau erabili **u**: http://goo.gl/ gunean. Itsatsi esteka eta **eskuratu** beste laburra.

b. **Idatzi** zure apple IDa.
Sartu zure apple IDa.

c. **Abiarazi** Skype.
Hasi Skype.

Lehen kodeketan hasieratik finkatutako aldaketa guztiak aplikatu dira. Hala ere, hauen eragina ikusteko ez da egokiena. Alde batetik itzulitako hitzen lemak aldatzen dira eta bestetik hitzen kasu markak. Emaitza orokorra aztertu ordez, bi partetan banatzea erabaki da, bata eta bestearen eragina ebaluatzeko.

6.2.2 Bigarren proba: lemak alde batetik, formemak bestetik

Lehen esperimentua nahiko orokorra izanik, egindako aldaketen eragina hobeto ikusteko banatzea erabaki da: rol semantikoek lemetan nola eragiten duten ikusi nahi da lehenik eta formemengan ondoren.

Horretarako, itzultzailearen exekuziorako aukera ezberdinak finkatu dira. Rolen tratamendua egin nahi den kasuetan lemak, formemak edo biak aldatu nahi diren aukera daiteke. Biak aldatzea, lehen proban bezala uztea litzateke. Bestela, bi lanetako bakarra aplikatzen da. Ondoko 6.3 taulan aldaketa bata eta bestearekin lortutako BLEU zenbakiak ikus daitezke, bi corpusetarako:

	formemak aldatuta	**lemak aldatuta**
batch2a	19.71	19.61
news	2.05	2.05

6.3 Taula: 2. esperimenturako BLEU konparaketak

BLEU zenbaki hauek eta aurrekoen artean, ez da ezberdintasun handirik ageri . Sorkuntzako arazoak konpondu gabe egonik, BLEU zenbakia, baita irteerako esaldiak ere ez dira oso esanguratsuak. Hala, rolen analisirik gabeko esaldiekin alderatuz, formemak bakarrik aldatzen diren esaldietan ezberdintasun bakarra ageri da, lehen emaitzetan jadanik azaltzen zena: *da* aditza, *u* bilakatzen da, bi corpusetan.

Lemak bakarrik tratatzean berriz, ezberdintasun gehiago ageri dira. BVI-ko aditzen arabera itzulpen batzuk aldatu egin dira. Beheko 6.4 taulan, esaldi zati batzuetako adibideak ageri dira, rolen analisiaren aurretik eta ondoren dauden lemekin:

Jatorrizko esaldi zatia	Rol gabeko irteera	Lemak aldatutako irteera
one would **like** to think that...	pentsatu **nahi** dena	pentsatu **gustatzen** dena
had even **thought** about	ere **uste** dena	ere **pentsatu** dena
speaking to	hitz	mintzatu
looking for	bila	begiratu
found that	ikusi da	aurkitu da.
is becoming	bihurtzen da	bilakatuko da

6.4 Taula: Lemak bakarrik aldatuta ikus daitezkeen ezberdintasunak

Orokorrean, emaitza batzuk okerragoak, besteak hobeak eta besteak berdin geratzen direla ondorioztatu da, batzuetan sinonimoak ematen baitira. Hala ere, esaldiak banaka aztertuz, egon behar ez luketen aldaketa batzuk ikusi dira. Izen batzuei lema aldatu zaie, aditz bilakatuz:

Jatorrizko esaldi zatia	Rol gabeko irteera	Lemak aldatutako irteera
Select or tap the + **sign**.	Hautatu edo sakatu, + **ikurra**.	Hautatu edo sakatu, + **sinatuak**
Make the desired **changes**.	Egin nahi **aldaketak**.	Egin nahi **aldatu**.
Some devices need to have a minimum **charge** to turn on.	Gailuak batzuk **karga** gutxieneko bat da behar dena aktibatu.	Gailuak batzuk **kobratu** gutxieneko bat da behar dena aktibatu.

6.5 Taula: Izenei egindako lema aldaketa adibideak

Goiko 6.5 taulako adibideetan, ingelesezko beltzeko hitzen forma aditza ere izan daiteke, baina ez da hala (*sign* adibidez). *Mate* etiketatzaileak ordea, izenak ere etiketatzen ditu. Proiektuan izenezko rol semantikoak ez dira landu eta beraz kontuan hartu ere ez. Treex-eko kodeak, nodoak *wild_dump* eremuan *srl* azpieremu bat duela ikustean, aditz moduan tratatu du, izena izanik ere. Arazo honi konponbide bat emateko asmoz, egokitzapen batzuk eta hirugarren proba bat egitea erabaki da.

6.2.3 Hirugarren proba: aditza izatearen murriztapena

Aurreko ataleko emaitzetan, aditzak ez diren hitzen lemak ere aldatzen direla ikusi da. Alta, proiektuan rol semantikoak aditzen inguruan bakarrik aztertzen dira eta eramandako lanketak izenen inguruan ez du zentzurik. BVI-n informazio baliokidea bilatzean, aditzei dagokiena bakarrik topatuko da. Beraz, kodea eguneratu eta murriztapen hau kontuan hartzea erabaki da.

Honela, prozesuaren hasieran ingelesezko nodoak *wild_dump* eremuan *srl* azpieremu bat duen begiratzen den bezala, nodo horren ezaugarrietan aditza den ala ez begiratzen da, ondorengo ekintza guztiak kasu horretan bakarrik aplikatuz:

if ($anodes[$i]–>get_iset('pos') eq 'verb') {...}

Beraz, hitz bati aldaketaren bat aplikatzeko, *srl* eremu bat izan behar du eta aditza izan behar da, besteak beste. Aurretik ikusitako adibideak honela gelditzen dira, murriztapen hau gehitu ondoren:

2. probako irteera	**3. probako irteera**
Hautatu edo sakatu, + **sinatuak**	Hautatu edo sakatu + **ikurra**.
Egin nahi **aldatu**.	Egin nahi **aldaketak**.
Gailuak batzuk **kobratu** gutxieneko bat da behar dena aktibatu.	Gailuak batzuk **karga** gutxieneko bat da behar dena aktibatu.

6.6 Taula: Nodoa aditza ote den begiratu aurreko eta ondorengo emaitzak

Antzeko esaldi batzuk zuzendu diren arren, BLEU zenbakietan ez da ezberdintasunik ageri. Ondoko 6.7 taulan, nodoa aditza ote den begiratu aurreko eta ondorengo BLEU-ak ikus daitezke, bi corpusetarako:

	2. proba	**3. proba**
batch2a	19.61	19.61
news	2.05	2.05

6.7 Taula: BLEU zenbakiak aditza izatearen murriztapena gehitu aurretik eta ondoren

Hasieratik ia aldatzen ez den BLEU zenbakiak informaziorik ematen ez duenez, berriro ere esaldiak TrEd-en laguntzaz aztertu dira eta bertan zuzendu ez diren adibide gehiago topatu dira. Kasu batzuetan Treex-ek hitzen ezaugarriak gaizki etiketatzen ditu, aditz etiketak emanez behar ez luketenei. Beraz, hirugarren esperimentu honetan gehitutako murriztapenak akats honengan eragina du. Hona adibide batzuk:

Jatorrizko esaldi zatia	**Rol gabeko irteera**	**Lemak aldatutako irteera**
Slide **left**, then tap on My Videos.	Irristatzea, gero sakatu Nere bideoak, **ezkerreko** dena.	Irristatzea, gero sakatu Nere bideoak, **utzi** dena.
Detecting lies, or "**lie** spotting,"is an essential skill for everyone to acquire	Gezurra edo **gezur** puntu, detektatu "" trebetasuna ezinbestez guztiak eskuratu da.	Gezurra edo **etzana** puntu, detektatu "" trebetasuna ezinbestez guztiak eskuratu da

6.8 Taula: Izenak aditz moduan tratatutako adibideak

Goiko 6.8 taulako adibideetan, sorkuntza arazoa dela eta itzulpenak txarrak direla agerian dago, hitzen ordena eta abar. Hala ere, beltzez dauden hitzen forma ez da egokia; *left* hitza kasu honetan ezkerrari dagokio eta ez *utzi* aditzaren iraganari. Kodean nodoa aditza den edo ez kontrolatuz ere, Treex-ek gaizki etiketatu baitu, prozedurarekin jarraitzen da, izenen lemak aditz gisara tratatuz.

Kasu honetan aipatutako arazoa, Treex edo *Mate* tresnek etiketatzea hobeto egin ezean, ezin da tratatu. Honen ondorioz akats hauetatik abiatuz ez da beste esperimenturik pentsatu. Hala ere, beste esaldi batzuetan ezohiko emaitzak aurkitu dira, formemei dagokienez. Honek laugarren proba egitea bultzatu du.

6.2.4 Laugarren proba: formema eguneraketarako murriztapena

Aurreko probetan kodean egindako aldaketek batez ere aditzen lemetan izan dute eragina. Horiez gain ordea, beste arazo batzuk aurkitu dira, formemekin lotuta. Aipatu bezala, Treex-ek sortutako t-zuhaitzetan, aditz nodoen umeak aztertzen dira hauei formema atributua aldatu edo ez erabakitzeko. Aztertzen diren ume horiei esleitutako argumentu zenbakia BVI-n agertzen direnekin bat baldin badator, formema atributua aldatzen zaio, hau da, kasu marka. Batzuetan, zuhaitz horietan aditz nodo baten azpian beste aditz bat egon daiteke, eta argumentu zenbakia bat etorri ere bai.

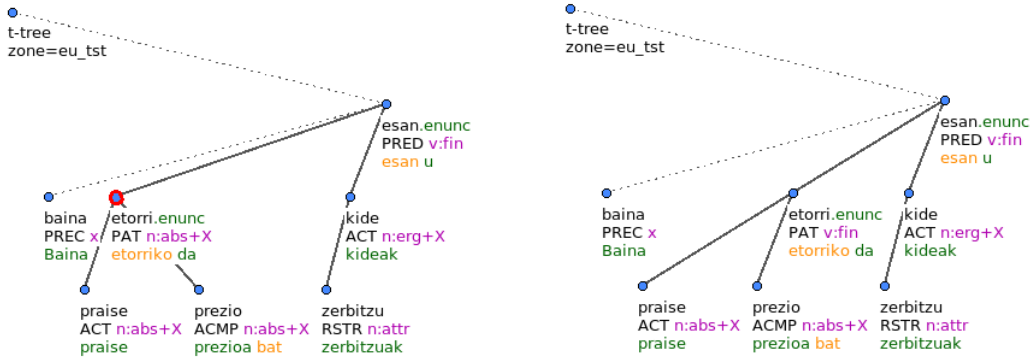
Kasu horretan, aditzari kasu marka aldatzen zaio, izena balitz bezala. Honek bi ondorio ekar ditzake: bata, nodo horren kasu marka okertzea eta bestea, nodo hori aditza ote den egiaztatzean ezezkotzat jotzea eta haren umeen formemak ez tratatazea.

Arazo honi aurre egiteko aurreko atalekoaren antzeko murriztapena gehitu da. Aditz nodo baten umeak aztertzen hasterakoan, umea aditza den edo ez begiratu behar da, hala bada ezer ez egiteko. Are gehiago, beste akatsak ere ekiditeko, ume hori izena izan dadin ziurtatzen da, formema jakin bakarra onartuz:

if ($child–>formeme =~/ ^n:.+/)

Aldaketa sinple honekin, itzulpena berriro egin eta irteera aztertu da. Beheko (16) adibideari dagozkion euskarazko t-zuhaitzak ageri dira ondoko 6.1 irudian. Murriztapena gehitu aurretik eta ondorengo zuhaitzak dira. Bertan *etorri* aditzaren formema *n:abs+X* izatetik *v:fin* izatera zuzentzen dela ikus daiteke.

(16) *But the praise comes with a price, service members say.*



6.1 Irudia: Guraso eta umea aditzak direnean egin den aldaketa

Ondoko 6.9 taulako BLEU zenbakiak lortzen dira, umea izena izatearen murriztapena gehitu aurretik eta ondoren:

	3. proba	**4. proba**
batch2a	19.61	19.61
news	2.05	2.05

6.9 Taula: BLEU zenbakiak, umea izena izatearen murriztapena gehitu aurretik eta ondoren

Beste behin, BLEU zenbakiak ez dira aldatu. Esaldi batzuk eskuz aztertu ondoren, aurreko probetan egin den bezala, ez da beste akats edo kontuan hartu gabeko punturik aurkitu. Beraz, metodoz aldatu eta ondoko atalean azaltzen den moduan egin da hurrengo ebaluazioa.

6.3 Eskuzko azterketa: ebaluazio kualitatiboa

Probak egin diren heinean, esaldi batzuk aztertu dira, inkongruentziak aurkituz. Honela, kodea pixkanaka egokitu da, aurreko ataleko etapak jarraituz. Azkenik lortutako itzulpenean ezohikotasun edo akatsik aurkitu ez eta emaitza egokitzat jo da.

Aipatu bezala, BLEU zenbakiak ez dira esanguratsuak sorkuntza arazoa dela eta. Lema edo formemak hobetu arren, emaitzetan ez da aldaketa hori islatzen. Horregatik, azterketa kualitatiboa egitea erabaki da; hau da, esaldi berriak zaharrekin konparatu, lema eta kasu marken bilakaera ikusi eta aldaketa horiek ebaluatu dira.

Eskuzko azterketa hau bi pausotan egin da: i) rol semantikoen eraginez aldaketa bat jasaten duten hitzen kontaketa egin da, adibide batzuk gordez eta ii) adibideak informazio

gehiagorekin gorde dira, ezaugarri aldaketak argiago ikusteko moduan, eta hizkuntzalarien laguntzaz aztertu dira.

6.3.1 Kopuruak eta adibide batzuk

Lehen pauso batean kontaketa berezi bat egin da. Nodo bakoitzeko eraman da kontaketa, irizpide batzuen arabera. Honako ezaugarriak dituzten nodoak kontatu dira, bost multzoetan banatuz:

- Jatorriko a-nodoko *wild_dump* eremuan *srl* eremua dutenak
- Hauen artetik aditzak direnak
- Proposatutako itzulpenak BVI-koekin bat egiten dutenak: oroitu, *srl* eremuan *Mate*k aditz adiera bat proposatzen duela, eta BVI-tik ingelesezko adiera horri dagozkion euskarazko aditz batzuk lortzen direla. Euskarazko nodoak esleitua duen itzulpena, BVI-ko ordain horietako bat baldin bada, egokia dela suposatu da eta ez zaio aldaketarik egin.
- Lema aldaketak: Aurreko kasua bete ezean, TectoMT-ren beste proposamenak aztertzen dira. Horietako bat BVI-ko bat baldin bada, unean esleitutako lema, horrekin ordezkatzen da.
- Formema aldaketak: aurreko puntuaren antzera, uneko nodoaren formema berez esleitua zuenaren ezberdin batekin ordezkatuz gero, hemen kontatzen da.

Esaldi eta nodo guztiak tratatu ondoren, emaitzak pantailaratu egin dira, zenbakizko datuak lehenik eta hauetarako gordetako adibideak jarraian, alfabetikoki ordenatuta. Honako datuak gorde dira adibideetan, kopuruez gain:

- Lema berdin uzten denean, ingelesezko adiera, euskarazko ordaina eta ingelesezko esaldia.
- Lema aldaketa bat gertatzean, ingelesezko adiera, euskaraz hasieran esleitutako lema, lema berria eta ingelesezko esaldia.
- BVI-n aurkitu ez diren ingelesa-euskara lema bikoteak.

TectoMT itzultzailea exekutatzerakoan eskuratu nahi diren adibideen kopurua zehazten da, nahiz eta kontaketa nodo guztiekin egin. Ondotik ikus daitezke *batch2a* eta *news* corpusetarako lortu diren kopuruak:

- *batch2a* corpusa:
  Etiketatutako nodoak: 5003

5003tik, hauek aditzak dira: 3182
BVI hiztegian aurkitu eta aldatu gabeko lemak: 1337
Lema aldatuko litzatekeen kopurua: 207
Formema aldatuko litzatekeen kopurua: 263

- *news* corpusa:
  Etiketatutako nodoak: 4948
  5003tik, hauek aditzak dira: 2562
  BVI hiztegian aurkitu eta aldatu gabeko lemak: 1188
  Lema aldatuko litzatekeen kopurua: 372
  Formema aldatuko litzatekeen kopurua: 686

Emaitzen lehen irakurketa kopuruekin egin da. Rolak etiketatuta dituzten nodoen artean aditz moduan tratatzen direnen kopurua nabarmen txikiagoa da. *batch2a*-ren kasuan %64 eta *news*-enean %52 besterik ez. Beste nodoak izenak edo etiketa txarrak dira, alde batetik *Mate*k gaizki etiketatutakoak izan daitezke edo TectoMT-k bestela. Tratatzen diren aditz nodo horietatik %42 (*batch2a*) eta %46-k (*news*) jadanik BVI-k proposatzen duen lema bera dute eta beraz, ez dute aldaketarik eragiten. Aldatzen direnak berriz %6.5 eta %14.5-a bakarrik dira. Falta diren 1638 eta 1002 aditz nodoetan ez da aldaketarik egon. Kasu hauetan, TectoMT-k proposatutako itzulpen aukeren artetik ez da BVI-n dagokion sarrerarik aurkitu ingelesa-euskara bikote horietarako, bietan bat gutxienez falta baita.

(17) *RSS is a technology that allows you to add information from different sources , for example if newspapers make available item , you can receive the latest news in your aggregator.*

Zenbaki hauek diotena ulertzeko gordetako adibideak begiratu dira. Goiko (17) adibidean, itzulpena egitean *aukera* hitza proposatzen zen eta *onartu*-rekin ordezkatu da, *allow.01* adierarako. BVI-n, adiera horretarako euskaraz *onartu* eta *utzi* proposatzen dira. TectoMT-ren itzulpenen artean *onartu* dagoenez, hori esleitu zaio. Aldaketa honek itzulpena hobetu duela esan daiteke.

Esaldi honen antzera, beste adibideak ere landu dira. BVI-ko aditzekin bat egiten duten gehienak oinarrizko hitzak dira, orokorrak, domeinu batekin zerikusirik ez dutenak. BVI-n aditz eta adiera asko falta dira eta TectoMT-k ez ditu beti zentzuzko proposamenak egiten.

Azterketa hau gehien bat lemei begirakoa da eta lehen iritzi bat lortzeko oso baliagarria da. Hala ere, benetan emaitza hobeak edo okerragoak diren jakiteko azterketa sakonago baten beharra dago. Jakinez predikatu batekin lotutako formema bat baino gehiago alda daitezkeela, aldatutako formema guztien kopurua ez da hain esanguratsua. Ezin da jakin zenbat predikaturen inguruan egiten den formema eguneraketa. Aditzaren lema lehen bezala utzi edo ez, *srl* eremuko argumentuak beti hartzen dira kontuan aditzaren umeen formemetarako. Beraz, 3182 eta 2562 aditzen ume guztiei aldatzea da posible, beti ere dagozkien murriztapenak betez gero.

Adibideetan ingelesezko esaldia bakarrik gordez, ezin da euskarazko esaldiko berrikuntzarik ikusi. Lema berria zein den jakinik ere, formemen eragina ez dago aztertzerik. Esan beharra dago hala ere sorkuntza arazoengatik hasieratik ikusi dela euskarazko esaldiek ez dutela informazio handirik emango. Horregatik, emaitzak beste modu batean aurkeztea erabaki da.

6.3.2 Eskuzko azterketa zehatzagoa

Itzulpenean aldatu diren eta ez diren lemak eta formema aldaketak kontatu ondoren, hauen ezaugarri zehatzak ikusi nahi izan dira. BVI-k proposatzen dituen kasu markak baliatuz, esaldiko egitura hobetu behar litzateke. Hala ere, aurreko probetan ezin izan da halakorik ikusi. Lemak ere hobetu behar lirateke.

Pausu honetan, alde batetik lema zahar eta berriak konparatu nahi izan dira, eta formema zahar eta berriak bestetik, predikatua eta kasu marka aldatu beharreko hitza bera zein diren kontuan harturik, baita bukaeran lortutako esaldia ere. Datu guzti hauek bi taulatan egituratu dira.

Ondoko 6.10 taulan ikus daiteke lemak aztertzeko taula egitura, *news* corpuseko adibideekin. Ingelesezko predikatuaren adiera, hasieran TectoMT-k euskararako esleitutako lema eta orain, rol semantikoen azterketaren ondoren esleitu zaiona ageri dira. Ondotik esaldi osoa ikus daiteke ingelesez, 3 zutabetan banatuta, erdiko zutabea predikatuari dagokiona izanik. Honela, errazago topatzen da esaldi luzeetan.

EN ad.	Lema zaharra	Lema berria	Esaldia		
look.01	bilatu	begiratu	Interrogators often falsely signal that an interview is over just to	look	for that post-interview relief.
turn.02	biratu	bihurtu	At worst, they accuse his administration of	turning	a blind eye.
join.01	sartu	elkartu	He asked if he could	join	her in her room for a drink.

6.10 Taula: *lema_azterketa_news.txt* fitxategi egitura

Bestalde, formemak aztertzeko beste taula bat sortu da, egitura konplexuagokoa. Ingelesezko predikatuaren adiera, dagokion euskarazko ordaina eta tratatua izaten ari den umea ageri dira. Ume horri dagozkion informazioak ikus daitezke: *Mate*k emandako argumentu zenbakia, dagokion ingelesezko hitzaren formema, euskarazko formema zaharra eta berria. Ondotik, ingelesezko esaldia ageri da. 6.11 taulan ikus daiteke:

EN ad.	EU Ad.	Umea	Arg	EN form.	EU form. zaharra	EU form. berria	Esaldia		
kill.01	hil	edari	A0	n:subj	n:abs+X	n:erg+X	The drink can	kill	the smell.
say.01	esan	polizia	A2	n:to+X	n:attr	n:dat+X	They put me into the psychiatric unit, and when I got out, I remember	saying	to my command officer, 'That was n't his trial'.
like.01	gustatu	#PersPron	A0	n:subj	n:abs+X	n:dat+X	We are not very showy, we do not	like	ostentation.

6.11 Taula: *formeme_azterketa_news.txt* fitxategi egitura

Bi taula hauekin, itzulpenak hobera egin duen, okerrera edo berdin geratu den jakin nahi da. Horretarako, *news* corpuseko emaitzen errenkadak ausaz ordenatu eta hizkuntzalari batzuek bakoitzeko lehen 100-ak aztertu dituzte, eskuz. Beraiek betetako taulek bi zutabe gehiago dituzte, eskatutakoaren emaitzak ongi adierazteko.

Lehen zutabean, *hobeto*, *okerrago* edo *berdin* hitzak agertzen dira. Berdin edo okerrago gelditzen diren kasuetan arrazoia zehazteko eskatu zaie, arrazoi posible batzuk proposatuz. Bigarren zutabean arrazoi hauei dagozkien laburdura batzuk idatzi dituzte, ondoko bi puntuetan azaltzen direnak.

- Lemetako akatsak sailkatzeko irizpideak:

    1. Ingelesezko adiera: Kode hau ingelesezko aditz adiera gaizki etiketatuta dagoenean jartzen da, *Mate*n akatsa denean alegia.
    2. BVIn landu gabea: akats hau domeinu arazo batekin lotu daiteke. Aurkitutako itzulpena orokorrean egokia izan daiteke, baina agian kasu horretarako behar litzatekeena ez dago BVI-n.

3. MT aukeraketa: TectoMT-ren lema proposamenetan egokiena ez den itzulpen bat lehenago proposatu eta BVI-n existituz gero, hori aukeratuko da.
4. Bestelakoa: aurreko arrazoi bat ere bete ez eta hala ere okertzen denean.

- Formemetako akatsak sailkatzeko irizpideak:

1. Ingelesezko adiera: Lemak aztertzean bezala, *Mate*k adiera txarra ematean jartzen da kode hau.
2. Argumentua: hau ere *Mate*n akats batekin lotuta dago, argumentu zenbaki desegokia esleitzeari baitagokio.
3. BVIn landu gabeko kasua: BVI-ko zerrendan argumentu bakoitzerako kasu ezberdinak proposatzen dira, gerta liteke argumentu jakin baterako zerrendan egokia den bakar bat ere ez egotea. Kasu horretan etiketa hau jartzen da.
4. MT aukeraketa: Argumentuaren kasu aukeretan egokia egon arren, posible da txarra aukeratzea. Oroitu BVI-ko egitura:

accept.01			
	onartu		
		A0	erg
		A1	abs
			konpl
			mod
			par
		A2	dat

Adibide honetan A1 argumenturako 4 aukera daude. Gerta daiteke TectoMT-k proposatutako formemak zerrenda honekin konparatzean desegokia den bat lehenago aurkitzea behar lukeena baino.

5. Egitura sintaktikoa: Orokorrean ongi dagoen kasu marka bat ez da esaldi guztietarako egokia. Egitura sintaktikoak BVI-n tratatu gabeko salbuespena eskatzea gerta liteke, pasiboen kasuan adibidez.
6. Bestelakoa: aurreko arrazoietako bat ere bete ez eta hala ere aldaketa aurrekoa baino okerragoa denean etiketatzen da honela.

Irizpide hauek jarraituz beraz, bi zutabe gehiagoko taulak berreskuratu dira. Egitura hauek azaltzeko erabili diren bi taulen irteera ondoko 6.10 eta 6.11tauletan ikus daiteke:

Ebaluazioa	Arrazoia	EN ad.	Lema zaharra	Lema berria
okerrago	2- BVIn langu gabea	look.01	bilatu	begiratu
		Interrogators often falsely signal that an interview is over just to **look** for that post-interview relief.		
berdin	1- ingeleseko adiera	turn.02	biratu	bihurtu
		At worst, they accuse his administration of **turning** a blind eye.		
hobeto	-	join.01	sartu	elkartu
		He asked if he could **join** her in her room for a drink.		

6.12 Taula: *lema_azterketa_news.txt* fitxategi egitura

Ebaluazioa	Arrazoia	EN aditza	EU Aditza	Umearen Lema	Arg	EN formema	EU formema zaharra	EU formema berria
hobeto	-	kill.01	hil	edari	A0	n:subj	n:abs+X	n:erg+X
		The drink can **kill** the smell.						
hobeto	-	say.01	esan	polizia	A2	n:to+X	n:attr	n:dat+X
		They put me into the psychiatric unit, and when I got out, remember **saying** to my command officer, "That wasn't his trial.						
hobeto	-	like.01	gustatu	#PersPron	A0	n:subj	n:abs+X	n:dat+X
		We are not very showy, we do not **like** ostentation.						

6.13 Taula: *formeme_azterketa_news.txt* fitxategi egitura

6.12 taulan hiru ebaluazio ezberdin atera dira. Adibidez, *look.01* aditzerako BVI-n *begiratu* bakarrik proposatzen da. Hasieran TectoMT-k *bilatu* esleitua zuen, behar den bezala, baina okertu egin da.

Ondoko 6.13 taulan ageri diren hiru adibideetan berriz hobekuntza bat egon da. Esaterako, *the drink can kill*-en itzulpena *edariak hil dezake* izango litzateke. *edari*-k, subjektuak, ergatibo marka du orain, behar den bezala. Lehen zegoen bezala absolutiboan utziz gero, *edaria hil dezake* itzul zitekeen.

Taula bakoitzeko, 100 esaldi aztertu dira modu honetan. Beheko 6.14 taulan bata eta besteko emaitzak bildu dira, okerrago, hobeto edo berdin dauden esaldien kopuruekin. 6.15 taulan berriz okertu edo berdin geratu diren esaldietarako arrazoi ezberdinen maiztasunak bildu dira. Hala ere, berdin dauden esaldi batzuetarako ez da arrazoirik definitu, beraz, kopuruek ez dute bat egiten. Oroitu, honakoak dira arrazoi ezberdinak:

Formemen azterketarako:

1. Ingelesezko adiera gaizki etiketatua
2. Argumentua gaizki etiketatua
3. BVIn landu gabea
4. MT aukeraketa txarra
5. Egitura sintaktiko berezia
6. Bestelakoa

Lemen azterketarako:

1. Ingelesezko adiera gaizki etiketatua
2. BVIn landu gabea
3. MT aukeraketa txarra
4. Bestelakoa

	Formema azterketa	Lema azterketa
hobeto	74	41
berdin	6	30
okerrago	20	29

6.14 Taula: Eskuzko azterketako emaitzak

Formema azterketa (20 okerrago+6 berdin)							
Arrazoia	1	2	3	4	5	6	Orotara
Kopurua	6	8	1	0	10	1	26
Lema azterketa (29 okerrago+12berdin)							
Arrazoia	1	2	3	4			Orotara
Kopurua	14	23	0	4			41

6.15 Taula: Okertzeko arrazoiak kopurutan

Orain arte emaitzak ikustea zaila zen. Lema batzuk begiratzean, sorkuntza arazoak kontuan hartuz, ez zuen ematen hobekuntza handirik zegoenik, lema hobeak aurkitzen baitziren baina baita okerragoak ere. Hala ere, eskuzko azterketak emaitza positiboak erakusten ditu. Esaldi gehiagorekin egitea komeni den arren, hemen aldatutako formemen %20 bakarrik okertzen dira eta lemen %29.

Lemei dagokienez hobetzen eta berdin geratzen direnen kopurua oso ezberdina ez den arren, formema aldaketetan %74 hobetzeak rol semantikoak erabiltzearen onurak argi uzten ditu.

Formemen eguneraketako okertze edo berdinketen arrazoi ezberdinak ere kontatu dira. Formemen kasuan, gehienetan egitura sintaktikoa da kasu markak txarrak izatearen kausa. Arrazoi nagusiak hobeto ulertzeko adibide batzuk aurki daitezke ondoko lerroetan:

- **Egitura sintaktiko berezia**:
  Esaldi hauek salbuespen bat edo beste tratamendu bat behar lukete.

  (18) *At Yale Law School, the Veterans " Legal Services **Clinic** is **preparing** a case against the four major military academies for allegedly fostering a misogynistic atmosphere.*

  (18) adibidean *prepare.01* predikaturako, *prestatu* euskaraz, *Clinic* hitza A0 argumentu gisara etiketatu da eta bere formemaren balioa *n:abs+X* izatetik *n:erg+X*

izatera igaro da. Kasu bakarra aurreikusi da, orokorrean ongi egon litekeena: *Klinikak prestatu du, prestatuko du, prestatzen du* etab. Hemen ordea *Klinika prestatzen ari da* litzateke esanahia eta egitura hori ez da kontuan hartzen.

- **Ingeleseko adiera gaizki etiketatua**:
  Beste arrazoi nagusiak *Mate* tresnak egindako etiketatze txarrekin lotuta daude. Maiz, adierak gaizki etiketatzen ditu.

  (19) *"It **used** to take me an hour to unload a delivery, now it takes me four hours", Yu said.*

  (19) adibidean esaterako, *use.01* adiera eman zaio aditzari, erabiltzearena alegia. Hemen ordea, ohiturari buruzkoa da, *use to* forman erabiltzen dena. Adiera txarra dela eta, kasu marka bat edo bestearekin ere gaizki dagoenez, ebaluazioan *berdin* etiketa jarri zaio esaldiari.

- **Argumentua gaizki etiketatua**:
  Mate tresnak, adierez gain, argumentu zenbakiak ere ez ditu beti ongi etiketatzen.

  (20) *Scalia and Thomas **dine** with healthcare law challengers as court **takes** case.*

  (20) adibidean esaterako *dine* hitza *take.01* predikatuaren A0 argumentu bezala etiketatu da. Hau hartzailearen paperari dagokio, esaldi honetan *court* dena.

Hemen ez da BVI-n landu gabeko adibiderik edo bestelako arrazoien adibiderik erakusten. Izan ere ez dira ia batere gertatu eta BVI-n ez egotearen arrazoia errazagoa da ulertzen adibiderik gabe ere. Lemei dagokienez, akats nagusiak BVIn falta den informazioagatik gertatu dira, baita *Mate*n etiketatzeengatik ere. Ez da adibiderik erakusten, ideia formemetako berdina baita.

7. KAPITULUA

Jarraipena eta Kontrola

Kapitulu honetan proiektua amaitu ondoren egindako lanaren ebaluazio orokorra aurkezten da. Hasieran planifikatutako denbora, kalitatea... benetan egindakoarekin alderatu eta ongi eta gaizki ateratakoa ebaluatuko da.

7.1 Proiektuaren garapena

Proiektu hasieran 408 ordu beharko zirela estimatu zen eta 430 izan dira beharrekoak. Ondoko 7.1 taulan ikus daitezke desbideraketa honen xehetasun gehiago. Kudeaketa eta defentsaren prestaketan ezik, beste ataletan desbideraketak egon dira.

Atazak	Estimatutako orduak	Egindako orduak	Desbideraketa (%)
Planifikazioa	30	20	-33.33
Ezagutzak eskuratzea	15	12	-20
Garapena	98	82	-16.32
Probak	130	154	18.46
Kudeaketa	25	25	0
Memoria	90	116.5	29.44
Defentsa	20	20	0
Orotara	408	429.5	5.27

7.1 Taula: Estimatu eta erabilitako orduen arteko desbideraketak

Planifikazioa egiten uste baino denbora gutxiago igaro da. Proiektuen kudeaketako

ikasgaian ikasitako kontzeptuak oroitu eta aplikatzea ez da oso zaila izan eta beraz denbora aurreztu da.

Hasieran aipatu da iazko udan jada TectoMT eta Perl lengoaiarekin lan egin zela eta horregatik otsailean proiektuarekin hastean berriz oroitu besterik ez zela egin beharko. Iaztik ikasitakoak lana azkartu egin du. Gainera, garapenarekin hastean pixkanaka gaitasun gehiago eskuratu dira, hasieran landu gabeak. Gauza bera gertatu da rol semantikoak, hauen etiketatzaileak eta BVI lexikoaren ezagutzarekin. Azken finean denbora pixka bat gutxiago pasa da tresnak aztertzen, garatzearekin batera sakonago landu baitira.

Garapenean ere estimatutakoa baino denbora gutxiago behar izan da, arazo handirik gabe lortu baita lehen bertsioa kodetzea eta errorerik gabe exekutatzea. Probak egiten hastean ikusi dira kodean arazo edo inkoherentziak, beraz, kodeari egindako aldaketak esperimentazio partean sartzen dira. Honengatik dago probetan honelako desbideraketa. Gainera, emaitzak aztertzen hastean, egokiak ez ziren emaitzak lortu arren, arrazoiak aurkitzea ez da beti erraza izan eta denbora asko igaro da horien bila.

Memoria idazteko ere estimatu baino denbora gehiago erabili da. TectoMT itzultzaile sistema ongi azaltzea uste baino zailagoa egin da eta denbora asko hartu da. Egindako probak ongi egituratu eta modu argian azaltzea ere ez da lan erraza. Azken finean, memoria idazteko denbora gehiago behar zen eta hasieran ez da guzti hori ongi pentsatu.

Orokorrean, proiektu hasiera uste baino azkarrago joan da. Emaitzen azterketa eta memoria idazterakoan denbora gehiago behar izan da, beraz oreka mantendu egin da.

7.2 Komunikazioa

Hasieran finkatu bezala, proiektu zuzendari batekin batera lan egin da proiektu osoan zehar, eguneroko aurrez-aurreko harremanean. E-mailak ez dira asko erabili, bilera batzuk adosteko besterik ez. Bi zuzendariekin bilerak egin dira, nahiko tarte erregularretan. Hasieran bilera gutxiago egin dira, bi edo hiru astez behin. Memoriaren idazketa eta esperimentazioaren zatia hobeto jarraitzeko berriz, astero egin dira bilerak.

7.3 Kalitatea

Atal honetan hasieran definitutako kalitate mailak bete direla ziurtatzen da.

- **Komunikazioa**: Hasieran pentsatu bezala, komunikazio erregularrak mantendu dira eta interesatuek proiektuaren berri izan dute momentu oro, beraien iritzia eta aholkuak emanez.
- **Produktua**: Kalitate maila minimoa betetzea bermatu da, TectoMT-n rol semantikoak kontuan hartuz lema eta kasu markak aldatzeko aukera baitago orain. Ebaluazio kuantitatiboak hobera egin ez duen arren (BLEU zenbakia) eskuzko azterketak hobekuntzak erakutsi ditu, beraz maila egokia lortu da. Hobeto ikusten diren emaitzak lortzeko denbora gehiago beharko litzateke, beste hainbat arlo baitaude lantzeko, ondorioetan azaltzen den moduan.
- **Memoria**: EHU-k ezarritako formatua jarraitu da. Zuzendariek hainbatetan idatzitakoa zuzendu eta aholkuak eman dituzte ulerterrazagoa izan dadin.
- **Defentsa**: Gardenkiak prestatzeko ere hasieran finkatutako baldintzak bete dira. Aurkezpena oraindik egin ez den arren beste pertsona batzuekin egin da eta beraz hori ere ahal bezainbat bermatua dago.

7.4 Arriskuak

Proiektu hasieran arriskutsu bezala kontsideratu diren puntuen bilana egin da hemen:

- Informazio galera: TectoMT sisteman egindako aldaketetan ez da inongo arazorik egon; ez da nahi gabeko akatsik egin eta IXA taldeko zerbitzariek ez dute arazorik izan. Hartutako neurriei esker ez da ezer galdu.
- Aurreikuspenak ez betetzea: aipatu moduan, gaizki planifikatutako ordu kopuruak egon dira. Bestetik, behin baino gehiagotan izan dira osasun arazo txiki batzuk. Hala ere, mugarrietarako aurreikusita zegoen denboragatik ez dute arazorik sortu.

8. KAPITULUA

Ondorioak eta etorkizunerako lana

8.1 Ondorioak

Dokumentu honetan azaldu den moduan, proiektuaren helburua TectoMT erregeletan oinarritutako itzultzaile automatikoari analisi semantikoa gehitzea izan da honen emaitzak hobetzen saiatzeko, ingelesetik euskarara. Horretarako, rol semantikoak erabili dira. Oroitu rol semantikoek esaldi bateko hitz bakoitzak esaldian jokatzen duen papera adierazten dutela, aditz edo ekintza nagusia kontuan hartuta.

Zehazki, proiektu hau burutzeko pauso ezberdinetan banatu da lana:

1. Rol semantikoak automatikoki etiketatzeko Mate tresna TectoMT-n integratu da.
2. BVI lexikoiak ingelesezko aditz adiera ezberdinetarako euskarazko ordain eta rolei buruzko informazioa ematen du. TectoMT eta gure beharretara moldatu behar izan da.
3. *Mate*n etiketak eta BVI-k emandako informazioa baliatu eta lotuz, itzulpenean erabiltzea lortu da.
4. Lortutako emaitzak aztertzeko esperimentazioa egin da, pixkanaka aldaketa batzuk gehituz.
5. Eskuzko azterketa prestatu da hizkuntzalarien laguntzaz egin eta aztertzeko.

Hala, adieraren araberako lemak jarri zaizkie hitzei, baita inguruko hitzen kasu markak aldatu ere (ikus 5 kapitulua).

Aurreko 6. kapituluan, egindako aldaketak aztertzeko emandako pausoak azaldu dira.

Horietan azkena hizkuntzalarien laguntzaz ezaugarri aldaketak ebaluatzea izan da, ondorio positiboak ateraz, %30 baino gutxiago baitira rol semantikoen erabileraren eraginez okertzen diren lemak, baita kasu markak ere.

Azterketa honi esker, egindako lana onuragarria dela ondorioztatu da. Hala ere, %30-ak okerrago egoten jarraitzen du. Bestetik, itzultzaileak oraingoz ez ditu euskarazko esaldiak ongi sortzen, beraz, eragindako aldaketak ez daude ikusgarri. Alderantziz, batzuetan itzulpenek okerragoak dirudite.

Orokorrean txarrak diren emaitzetarako, hainbat arrazoi nabarmendu daitezke:

- BVI-ko aditz kopurua: izan euskaraz edo ingelesez, adiera asko falta dira. Gehienetan kasu bakarra tratatzen da. Honek aldaketarik ez egotea bultzatzen du, edo egotekotan okerrera egitea.
- Mate-n etiketak: izenezko rolen etiketatzea egiten duenez nahaste batzuk gertatu daitezke, nahiz eta ahal den moduan hori tratatzeko saiakera egin den. Horrez gain adiera edo argumentu zenbakiak ez ditu beti ongi etiketatzen.
- TectoMT-ren etiketak: Mate-n antzera, TectoMT-k ere batzuetan etiketa txarrak jartzen ditu, aditzak ez diren hitzak aditz moduan tratatuz adibidez. Beste akats batzuk saihesteko aditz izatearen murriztapena gehituta ere, arrazoi honengatik ezin dira denak konpondu.
- Euskara-ingelesa baliokidetza fitxategia: Oroitu euskarazko eta ingeleseko PropBank-eko argumentu zenbakiak ez datozela beti bat. Horregatik BVI fitxategia moldatzerakoan, argumentu baliokidetza fitxategi bat erabili da. Hala ere fitxategi honetan ez daude BVI-ko aditz guztiak. Gainera anbiguotasunak ere sortzen dira, aukera bat baino gehiago proposatuz. BVI-ren tamaina eskasa izateari baliokidetza fitxategia ere txikia izatea gehituz, itzulpen onak aukeratzea asko murrizten da.

Arazo hauek, ikerketak jarraipen beharra duela erakusten dute. Emaitza kopuru txiki bat aztertuz hobekuntzak aurreikusi arren, muga hauek existitzen diren bitartean ez da aurreratzea lortuko. Horregatik, gradu amaierako proiektu moduan ikerketa urrunago eramango ez den arren, agian IXA taldeak lanean jarraituko du.

Dena den, esan daiteke proiektu honek zuen helburua bete dela. TectoMT itzultzailean rol semantikoen eragina kontuan har daiteke ingelesetik euskararako itzulpenean. Helburua hobekuntza bat ekartzen duen ala ez ikustea zen. Ebaluazio kuantitaboari dagokionez,

oraingoz emaitza berdintsuak lortzen diren arren, eskuzko azterketak hobekuntza argiak erakutsi ditu. Beraz, BLEU zenbakia hobetzeko aukera ere badagoela esan daiteke.

Maila pertsonalean ere proiektua oso aberasgarria izan da. Iazko udan IXA taldean egindako praktikengandik haratago, hobeto ezagutu ahal izan dut informatika eta hizkuntzalaritza batzen dituen arloa. Oso interesgarria iruditu zait, proiektu gisara zein gerora lan munduan jorratzeko ere.

8.2 Etorkizunerako lana

Etorkizuneko lana definitzen hasteko hainbat aukera dauden arren, zerrendatu berri diren akatsak har daitezke oinarritzat.

- BVI-ko aditz kopurua: lexikoia hedatzeak egindako lana aldatu gabe hobekuntzak ekarri behar lituzke, ikusi den moduan hobekuntzak baitaude baina kopuru txikian. Euskara zein ingelesezko aditz adiera gehiago sartuz, hauen kasu markekin batera, TectoMT-ko zerrendekin konparatzean berdinketa gehiago topatuko lirateke. Lexikoia handitzeko lana hizkuntzalarien laguntzaz burutu beharko litzateke.

- Mate-n etiketak: tresnak egiten dituen akatsak zuzentzea guretzat ezinezkoa da. Hala ere, argi dago garatzaileek hobekuntza bat egingo balute interesgarria litzakeela bertsio berria integratzea TectoMT-n.

  Horrekin lotuta, izenezko rol semantikoak eta euren eragina ere aztertu behar lirateke. Itzulpena hobetzeko onuragarriak izan daitezkeela pentsatzen bada, Mate-k jada etiketatzea egiten duenez, hori kontuan hartzeak ez luke lan handirik emango agian.

  Beste aukera bat ere etiketatzaile ezberdin bat erabiltzea da. Proiektu honetan ezagutu den ClearNLP esaterako egokia izan daiteke, nahiz eta hasiera batean Mate hobea zela iruditu. Beste etiketatzaileren bat ere aurkitzeko aukera egon daiteke.

- TectoMT-ren etiketak: honen akatsak zuzentzea ere agian ez da hain erraza, baina proiektua IXA taldean garatzen den heinean, etiketa horiek nola egiten diren aztertu liteke akatsen jatorria aurkitzeko.

- Euskara-ingelesa baliokidetza fitxategia: BVI lexikoiaren antzera, hemen baliokidetza gehiago sartu behar lirateke eraginkorragoa izan dadin. Honetarako ere hizkuntzalarien laguntza beharko da.

Puntu hauez gain, proiektuan zehar hainbatetan aipatu den sorkuntza arazoa dago. Euskarazko esaldien egitura ez da ongi sortzen, aditz asko ez dira jokatzen eta hitzen forma ere ez da beti egokia. Horren konponketa lana martxan dagoen arren, proiektu honekiko garrantzitsua da. Era honetan, ezaugarrietan ikusi diren emaitzak irteerako testuan ikusi ahal izango dira.

Azkenik, nahiz eta ideia bila luzaz jarrai daitekeen, lan berdina beste hizkuntza bikote batzuetara hedatzeak proiektu bat baino gehiago sor ditzake. Hau oinarritzat hartuz eta etiketatzeko existitzen diren aukeren arabera lan erraz edo zailagoa izan daiteke, Mate-k berak ingelesaz gain beste hizkuntza batzuetarako etiketatzea ahalbidetzen baitu.

Jarraipenerako lanaren abiapuntu moduan, BVI handitzearen ideia hartu da, eskuzko azterketaz gain beste fitxategi batzuk prestatuz. Fitxategi hauetan, aditzen agerpen maiztasunak kontatu eta handienetik txikienera ordenatu dira, BVI-n aurkitu diren ala ez kontuan hartu gabe. Honen helburua, batez ere informatikarekin lotutako domeinuko aditzak BVI-n gehitzea da. Aipatu bezala BVI-n adiera gutxi daude eta informatikakoak zehatzagoak direnez gutxitan aurkitzen dira egokiak. Hala ere *news* corpuserako ere kontatu dira maiztasunak.

Hona maiztasun fitxategi horien zati bat:

8.1 Taula: *batch2a* corpuseko maiztasunak

Aditza EN	**Maiztasuna**
click.01	503
go.02	142
tap.03	139
press.01	73
connect.01	56
want.01	37
go.10	26
remove.01	26
install.01	24
access.01	24

8.2 Taula: *news* corpuseko maiztasunak

Aditza EN	**Maiztasuna**
get.01	18
have.03	18
look.01	15
face.01	15
describe.01	12
bore.02	12
ask.01	12
run.01	11
think.01	10
become.01	10

Bibliografia

[Ahsan et al., 2010] Ahsan A., Kolachina P., Kolachina S., Sharma D., Sangal R. 2010. *Coupling Statistical Machine Translation with Rule-based Transfer and Generation*

[Aldezabal et al., 2013] Aldezabal I., Aranzabe M.J., Diaz de Ilarraza A., Estarrona A. 2013. *A methodology for the semiautomatic annotation of EPEC-ROLSEM, a basque corpus labeled at predicate level following the Propbank-Verbnet model*

[Astigarraga et al., 2009] Astigarraga A., Gojenola K., Sarasola K., and Soroa, A. 2009. *TAPE Testu-analisirako PERL erremintak.*

[Björkelund et al., 2009] Björkelund A., Hafdell L., Nugues P. 2009. *Multilingual semantic role labeling. In Proceedings of The Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), pages 43–48, Boulder*

[Čmejrek et al., 2004] Čmejrek M., Cuřín J., Havelka J. 2004. *Prague Czech-English Dependency Treeban: Any Hopes for a Common Annotation Scheme?*

[Estarrona, 2014] Estarrona A. 2014. *EPEC corpusa predikatu-mailan etiketatzeko oinarriak: EPEC-RolSem, BVI eta e-ROLda*

[Estarrona et al., 2015] Estarrona A., Aldezabal I., Díaz de Ilarraza A. eta Aranzabe M.J. 2015. *Methodology for the semiautomatic annotation of EPEC-RolSem, a Basque corpus labelled at predicate level following the PropBank/Verbnet model*

[Fillmore , 1968] Fillmore Charles J. 1968. *The case for case* in E. Bach and R. Harms, eds., *Universals in Linguistic Theory*, Holt, Rinehart and Winston, New York.

[Filmont, 2014] Filmont C. 2014. *Les rôles sémantiques: Théories et applications*

[Hutchins et al., 1992] Hutchins W.J., Somers H.L. 1992. *An Introduction to Machine Translation*

[Jayaraman et al., 2005] Jayaraman S., Lavie A. 2005. *Multi-Engine Machine Translation Guided by Explicit Word Matching*

[Kay et al., 1992] Kay M., Gawron J.M., Norvig P. 1992. *Verbmobil: A Translation System for Face-to-Face Dialog*

[Kipper, 2005] Kipper K., 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*

[Labaka et al., 2015] Labaka G., Jauregi O., Diaz de Ilarraza A., Ustaszewski M, Aranberri N., Agirre E. 2015. *Deep-syntax TectoMT for English-Spanish MT*

[Laparra, 2015] Laparra E. 2015. *Implicit semantic roles in discourse*.

[Leja et al., 1998] Leja C., Malaviya A., Peters L. 1998. *A fuzzy statistical rule generation method for handwriting recognition*

[Palmer et al.,] Palmer M., Kingsbury P., Gildea D. 2005. *The Proposition Bank: An Annotated Corpus of Semantic Roles*

[Papineni et al., 2002] Papineni K., Roukos S., Ward T., Zhu W. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*

[Puerto, 2016] Puerto A. 2016. *TectoMT bidezko euskarazko itzulpenen erroreen detekzio eta zuzenketa*

[Rosa et al., 2015] Rosa R., Dušek O., Novák M., Popel M. 2015. *Translation Model Interpolation for Domain Adaptation in TectoMT*

[Žabokrtský et al., 2008] Žabokrtský Z., Ptáček J., and Pajas P. 2008. *TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer*

[Zapirain, 2010] Zapirain B. 2010. *Rol Semantikoen Etiketatze Automatikoa: Rol Multzoak eta Hautapen Murriztapenak*

[1] *CoNLL 2009 Shared Task*-eko web orria: http://ufal.mff.cuni.cz/conll2009-st/

[2] *ClearNLP* tresnaren webgunea: https://clearnlp.wikispaces.com/

[3] *Mate* tresnaren webgunea: https://code.google.com/archive/p/mate-tools/

[4] Pragako unibertsitateko TectoMT-ren azalpenak : http://ufal.mff.cuni.cz/tectomt

[5] Pragako unibertsitateko Treex-en azalpenak : http://ufal.mff.cuni.cz/treex

[6] *Prague Dependency Treebank*-en azalpena: http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/Doc/whatis.html

[7] *Propbank*-eko aditz zerrenda: http://verbs.colorado.edu/propbank/framesets-english-aliases/

[8] *QTLeap* proiektuko helburuak: http://qtleap.eu/goals/

[9] Sareko Euskal Gramatikako Hizkuntzaren prozesamenduari buruzko atala: http://www.ehu.eus/seg/hizk/1

[10] *Verbnet* azalpena: http://verbs.colorado.edu/~mpalmer/projects/verbnet.html