EKAIA (2019). DOI: 10.1387/ekaia.19745

# Dependentzia Unibertsalen eredura egokitutako euskarazko zuhaitz-bankua

*(Universal Dependencies based Basque Treebank)*

*Maria Jesus Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Díaz de Ilarraza, Iakes Goenaga, Koldo Gojenola, Larraitz Uria*

Ixa Taldea, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU)

maxux.aranzabe@ehu.eus

*Jasoa: 2018-05-21*

*Onartua: 2018-09-27*

**Laburpena:** *Hizkuntzaren Prozesamenduan kokatzen den Dependentzia Unibertsalen proiektuaren helburua da hainbat hizkuntzatan sortu diren dependentzia-ereduan oinarritutako zuhaitz-bankuak etiketatze-eskema estandar berera egokitzea. Artikulu honetan, eredu horretara automatikoki egokitu den euskarazko zuhaitz-bankua aurkezten da. Egokitzapen-lan hori nola gauzatu den deskribatzen da eta, horretan oinarrituta, jatorrizko eta Dependentzia Unibertsalen eredura egokitutako zuhaitz-bankuen antzekotasunak eta desberdintasunak ere azaltzen dira.*

**Hitz gakoak:** Hizkuntzaren Prozesamendua, Dependentzia Unibertsalak, zuhaitz-bankua, sintaxia.

**Abstract: In the Natural Language Processing research area, the aim of the Universal Dependencies project is to convert dependency based treebanks developed in different languages into the same standard tagging scheme. This article presents the automatic conversion of the previously existing Basque treebank into this universal tagging scheme. This work describes how the conversion process has been carried out and highlights the similarities and differences between the original Basque treebank and the Universal Dependency based version of it.**

**Keywords:** Natural Language Processing, Universal Dependencies, treebank, syntax.

## 1. SARRERA

Dependentzia Unibertsalen (DU; ingelesez, Universal Dependencies, UD) proiektuaren[1] helburua da gramatika-erlazio unibertsalak proposatzea dependentzia-erlazioan dauden hitzak etiketatzeko hizkuntza edozein izanda ere [1]. Hizkuntzaren Prozesamenduan (HP), dependentzia-ereduan oinarritutako hainbat hizkuntzatako zuhaitz-bankuak modu berean etiketatuta izateak aukera emango du batetik, hizkuntza askotan erabil daitezkeen analizatzaile sintaktiko estatistikoak garatzeko eta, bestetik, hizkuntzen tipologiaren araberako egitura sintaktikoak aztertzeko. Lehen aukerari dagokionez, analizatzaile sintaktikoek etiketatze-eskema bakarrarekin lan egin ahal izango dute hizkuntzetako esaldien egitura sintaktikoak ikasten dituzten bitartean, eta bigarren aukeran, esaterako, bi hizkuntzen arteko antzekotasun sintaktikoak azter daitezke analizatzaile sintaktikoa hizkuntza jakin baten egitura sintaktikoekin entrena daitekeelako eta ikasitakoa beste hizkuntza bati aplikatu.

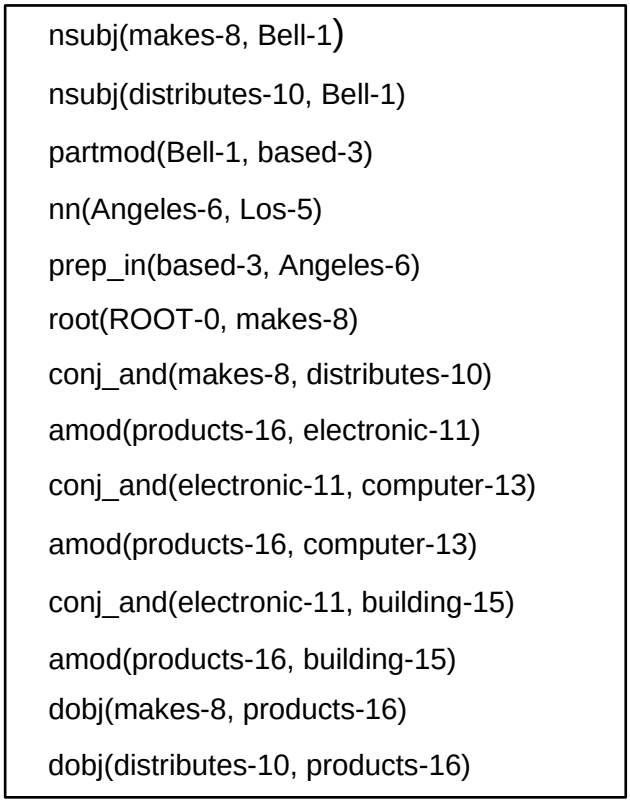
2008az geroztik, hizkuntza gehienentzat baliagarriak izango diren morfologiaren eta dependentzia-sintaxiaren etiketatze-eskema komunak bateratzeko hainbat saiakera egin dira: Google Universal Part-of-Speech Tags [2], HamleDT (Harmonized Multi-Language Dependency Treebank) [3, 4] eta Stanford Dependencies (SD) [5, 6]. Saiakera horien guztien emaitza da DU proiektua, “de facto”-zko estandarra bihurtu dena 2014tik aurrera.

DUetan erabiltzen den etiketatze-eskema Stanfordeko Dependentzietan (SD) [5, 6] oinarritzen da. SDak testuko elementu lexikoen arteko erlazioak modu errazean eta eraginkorrean erauzi nahi dituen erabiltzaileari begira diseinatuta daude. Horrela, bada, SDak hirukoteak dira: erlazioaren izena, gobernatzailea eta mendekoa. Esaterako, (1) esaldian [5] hirukoteetako bat osatzen dute *makes* aditzak, *Bell* izenak eta bien arteko erlazioa adierazten duen *nsubj* etiketak; bi hitzen arteko erlazio hori honela ulertu behar da: gobernatzailea den *makes* aditzaren subjektua da mendekoa den *Bell* izena.

(1) *Bell, based in Los Angeles, makes and distributes electronic, computer and building products.*

[1] http://universaldependencies.org/

Esaldi osoa 1. irudian adierazitako moduan etiketatu[2] da SD ereduari jarraituz.



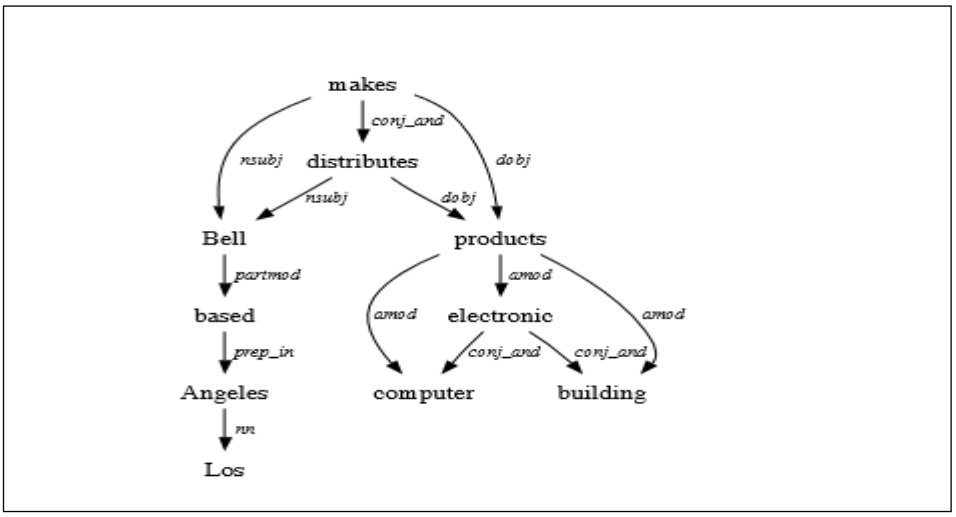

**1. irudia**. *Bell, based in Los Angeles, makes and distributes electronic, computer and building products* esaldiaren dependentzia-analisia.

Bi hitzen arteko erlazioa adierazteko erabiltzen den dependentzia-etiketa horietako lehen eremuan, bi hitzetatik gobernatzailea zein den eta esaldiko zenbatgarren hitza edo tokena[3] den esaten da, eta bigarren eremuan, berriz, mendekoa eta esaldian duen posizioa erakusten duen zenbakia. Horrela, dependentzia-etiketa horiek baliatuta esaldia osatzen duten elementu lexikoen arteko binakako erlazioak gauzatzen dira esaldiaren dependentzia-zuhaitza (2. irudia) lortuz. Dependentzia-

[2] Hemen alfabetikoki zerrendatuta ageri dira 1. irudiko dependentzia-etiketak beren esanahiarekin: amod (*adjectival modifier*): adjektibo modifikatzailea; conj_and (*conjunc*): juntagailua; dobj (*direct object*): objektu zuzena; nn (*noun compound modifier*): izen modifikatzailea; hau da, izen sintagmako izen kategoriako burua modifikatzen duen beste izen bat; nsubj (*nominal subject*): sintagma mailako subjektua; partmod (*non-finite verbal modifier that are participal in form*): partizipio modifikatzailea; prep_in (*prepositional modifier*): preposizioa; root (*root*): erroa.

[3] Tokenak hitzak ez ezik puntuazio-markak, zenbakiak, laburtzapenak edo antzeko beste edozein karaktere hartzen ditu bere barruan.

zuhaitzetan, gezien noranzkoak adierazten du erlazioan dauden bi hitz horietatik gobernatzailea dela geziaren abiapuntua den hitza eta mendekoa, berriz, geziaren helmuga den hitza.




**2. irudia**. *Bell, based in Los Angeles, makes and distributes electronic, computer and building products* esaldiaren irudikapen grafikoa SDak baliatuta [5].

Orain artean, 60 hizkuntzek hartu dute parte DUen 2.1 bertsioan eta 102 zuhaitz-banku daude erabilgarri [7]. Beren dependentzietan oinarritutako zuhaitz-bankuak eredu horretara automatikoki edo eskuz egokitu dituzten 60 hizkuntzak hauek dira: afrikaansa, alemana, antzinako greziera (1453ra arte), arabiera, bielorrusiera, bulgariera, daniera, errumaniera, Errusiako buriatera, errusiera, eslaviera, eslovakiera, esloveniera, estoniera, euskara, finlandiera, frantsesa, galiziera, gaztelania, goi-sorabiera, gotikoa, greziera modernoa (1453-), iparraldeko kurduera, iparraldeko samiera, hebreera, hindia, hungariera, indonesiera, ingelesa, irlandera, italiera, japoniera, katalana, kazakhiera, koptoera, koreera, kroaziera, latina, letoniera, lituaniera, marathera, nederlandera, norvegiera, persiera, poloniera, portugalera, sanskritoa, serbiera, suediera, suediar zeinu-hizkuntza, tamilera, telugua, turkiera, txekiera, txinera, ukrainera, uigurrera, urdua, vietnamera eta yue txinera.

Hizkuntza horietako bakoitzean egokitu edo sortu den zuhaitz-bankuaren ezaugarriak (tamaina, dokumentazio-lana, bertsioa, testu-motak…) proiektuaren webean kontsulta daitezke. Zenbait hizkuntzatan zuhaitz-banku bat baino gehiago

daude erakunde desberdinek sortu dituztelako; ondorioz, zuhaitz-banku horien tamaina eta edukia ere desberdina da.

Artikulu honetan, Ixa ikerketa-taldean[4] DUen eredura egokitu dugun euskarazko zuhaitz-bankua aurkeztuko dugu. Horretarako, sarrera honen ondotik, 2. atalean egokitzapen-prozesua nola gauzatu den azalduko dugu; zehazki, 2.1 azpiatalean abiapuntutzat hartu dugun zuhaitz-bankua deskribatuko dugu, 2.2 azpiatalean egokitzapena zein hiru mailatan eta nola gauzatu den azalduko dugu eta, horretan oinarrituta, bi zuhaitz-bankuen antzekotasunak eta desberdintasunak adieraziko ditugu, eta 2.3 azpiatalean egokitzapen-lanaren emaitza emango dugu ezagutzera. Amaitzeko, 3. atalean atera ditugun ondorioak zein diren esango dugu.

## 2. DEPENDENTZIA UNIBERTSALEN EREDUAN OINARRITUTAKO EUSKARAZKO ZUHAITZ-BANKUA

Dependentzia Unibertsalen proiektuan definitutako kategoria gramatikalen zerrenda unibertsalari eta gidalerroei jarraituta, zuhaitz-bankuen egokitzapena tokenizazio, morfologia eta sintaxi mailetan egin behar da CoNLL-U formatua erabiliz. CoNLL-U formatua berrikusi den CoNLL-X formatua da eta horren adibide bat ikus dezakegu 3. irudian; irudi horretan ikus daitekeen moduan, zutabeka antolatutako formatua da, aurrerago xehetasun handiagoarekin deskribatuko dena. CoNLL-X formatua deritzo 2006ko X. Computational Natural Language Learning (CoNLL) lehiaketan, dependentzietan oinarritutako analizatzaile sintaktiko-estatistikoak garatzeko erabili zen formatuari. Hain zuzen ere, CoNLL-X formatuan dugun euskarazko zuhaitz-bankua da egokitu duguna eredu horretara.

### 2.1. Abiapuntua

EPEC-DEP zuhaitz-bankua [8] Dependentzia Gramatikari jarraituz sintaktikoki etiketatu den 300.000 hitzeko Euskararen Prozesamendurako Erreferentzia Corpusa (EPEC) [9] da. DUen proiekturako, CoNLL-X formatuan dauden euskarazko EPEC-DEP zuhaitz-bankuaren 150.000 hitz [10] hartu dira gutxi gorabehera. Aurrerantzean, jatorrizko zuhaitz-bankua esango diogu 150.000 hitzez osatutako zuhaitz-banku horri.

[4] http://www.ixa.eus/

Jatorrizko zuhaitz-banku horretatik hartu den 3. irudiko *Bestetik, azken jardunaldietan presioa ukan dugu bizkar gainean, eta beste behin, hala izango da;* esaldiaren analisiak erakusten du zein den orain CoNLL-U izena duen formatuaren itxura. Irudi horretan ikus daitekeen moduan, esaldiko hitz bakoitzaren informazioa lerro batean jartzen da. Zutabeetan ageri den informazioa, berriz, hau da: lehen zutabean (ID, *identifier*), esaldia zenbat hitzek edo tokenek osatzen duten eta zein ordenatan ageri diren adierazi da 1etik 16ra bitarteko zenbakien bitartez; bigarren zutabean (WORD), esaldia osatzen duten hitzen formak ageri dira; hirugarrenean (LEM, *lemma*), hitz bakoitzari dagokion lema; laugarrenean (CPOS, *coarse part-of-speech*) eta bosgarrenean (POS, *part-of-speech*), hitz bakoitzaren kategoria eta azpikategoria, hurrenez hurren; seigarrenean (FEATS), ezaugarri morfosintaktikoak; zazpigarrenean (HEAD), esaldiko zenbatgarren hitzaren mende dagoen edo zein den hitzaren gobernatzailea eta zortzigarrenean (DEP, *dependency*) zein dependentzia-erlazio duen bere gobernatzailearekiko.

| ID | WORD | LEM | CPOS | POS | FEATS | HEAD | DEP |
|---|---|---|---|---|---|---|---|
| 1 | Bestetik | beste | DET | DZG | KAS:ABL\|NUM:S\|MUG:M | 0 | lotat |
| 2 | , | , | PUNT_MARKA | | PUNT_KOMA _ | 1 | PUNC |
| 3 | azken | azken | DET | ORD | _ | 4 | ncmod |
| 4 | jardunaldietan | jardunaldi | IZE | ARR | BIZ:-\|KAS:INE\|NUM:P\|MUG:M | 6 | ncmod |
| 5 | presioa | presio | IZE | ARR | BIZ:-\|KAS:ABS\|NUM:S\|MUG:M | 6 | ncobj |
| 6 | ukan | ukan | ADI | SIN | ADM:PART\|ASP:BURU | 10 | lot |
| 7 | dugu | *edun | ADL | ADL | MDN:A1\|NOR:HURA\|NORK:GUK | 6 | auxmod |
| 8 | bizkar_gainean | bizkar | IZE | ARR | BIZ:-\|POS:POSgainean\|POS:+\|KAS:INE | 6 | ncmod |
| 9 | , | , | PUNT_MARKA | | PUNT_KOMA _ | 8 | PUNC |
| 10 | eta | eta | LOT | JNT | ERL:EMEN | 0 | ROOT |
| 11 | beste_behin | beste_behin | | | ADB ARR MW:B | 14 | ncmod |
| 12 | , | , | PUNT_MARKA | | PUNT_KOMA _ | 11 | PUNC |
| 13 | hala | hala | ADB | ARR | _ | 14 | ncmod |
| 14 | izango | izan | ADI | SIN | ADM:PART\|ASP:GERO | 10 | lot |
| 15 | da | izan | ADL | ADL | MDN:A1\|NOR:HURA | 14 | auxmod |
| 16 | ; | ; | PUNT_MARKA | | PUNT_PUNT_KOMA | 14 | PUNC |

**3. irudia.** Jatorrizko zuhaitz-bankuko *Bestetik, azken jardunaldietan presioa ukan dugu bizkar gainean, eta beste behin, hala izango da;* esaldiaren dependentzia-zuhaitza CoNLL-U formatuan.

Euskararen kasuan, ezaugarri morfosintaktikoak dituen zutabeak (FEATS) informazio ugari bil dezake: kasu-marka, numeroa, mugatasuna (mugatua ala mugagabea den), aspektua, eta abar.

Behin DUen eredua aztertuta eta egokituko den corpusaren tamaina erabakita, egokitzapen-prozesu horretan kontuan hartu ditugun bi irizpide nagusiak hauek dira: i) esaldien egokitzapena automatikoki egingo da ahal den denbora eta eskulan gutxien inplikatzeko, eta ii) egokitutako esaldi horiek zuzenak izatea da helburua; hori dela

eta, zalantzazko kasuak baztertu egingo dira eta ziurtasun handiarekin ondo dauden esaldiak baino ez dira egokituko lehen urrats honetan.

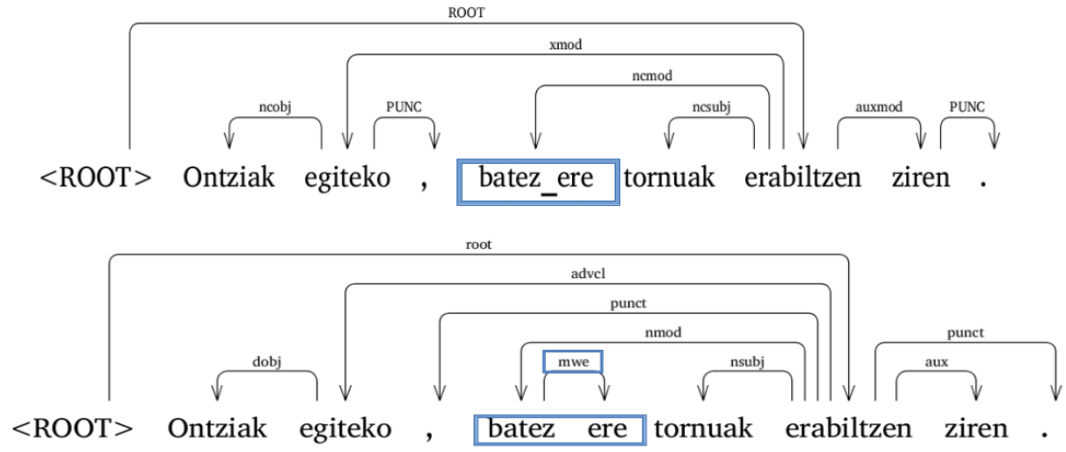
Esaldi horien mailakako egokitzapena azalduko dugu ondoren, hau da, tokenizazio, morfologia eta sintaxi mailetako egokitzapenak.

### 2.2. Mailakako egokitzapena

DUen etiketatze-lana sintaxiaren ikuspegi lexikalistan oinarritzen da; hau da, dependentzia-erlazioak hitzen artean gauzatzen dira. Horrenbestez, ezaugarri morfologikoak hitzen ezaugarri modura kodetuta daude eta ez da hitzak morfemetan banatzeko saiakerarik egin. Horrela, bada, garrantzia du nabarmentzeak etiketatzeko oinarrizko unitatea hitz sintaktikoa dela (ez hitz fonologikoak edo morfologikoak).

#### 2.2.1. *Tokenizazioa*

Tokenizazio mailan, analisi morfologikoan sarrera gisa erabiliko diren unitateak bereizten dira. DU proiektuan unitate horiek testu-hitzak dira. Testu-hitzak HPan, zuriunetik zuriunera bitarteko alfabetoko karakteren segidatzat hartzen diren testu-elementuak edo tokenak dira; hau da, hitzak ez ezik puntuazio-markak, zenbakiak, laburtzapenak edo antzeko beste edozein karaktere hartzen ditu bere barruan. Jatorrizko zuhaitz-bankuan ere hitzak dira aztergai; dena dela, euskararen kasuan, ziurtzat eta unitate bakartzat hartzen diren Hitz Anitzeko Unitate Lexikalak (HAUL) eta postposizio konplexuak [11] banatu behar izan dira eredura egokitzeko. HAUL ziurra da, esate baterako, “batez ere” adberbioa, “batez” eta “ere” osagaiek esaldi batean hurrenkera eta forma horiexetan agertzen diren guzti-guztietan hitz anitzeko adberbioaren interpretazioa izango dutelako; “menditik zehar” bezalako postposizio konplexuak, berriz, unitate bakartzat hartu dira bi osagaiko postposizio-sintagma horietan, postposizio beregainak (*zehar*) beren osagarriaren (*menditik*) atzetik erantsi gabe doazelako eta HAULetan bezala postposizio konplexua osatzen duten osagaiak hurrenkera horretan agertzen direlako. Banaketa horren adibidea da 4. irudiko *Ontziak egiteko, batez ere tornuak erabiltzen ziren.* esaldian ageri den *batez_ere* HAULa; HAUL hori hitz bakar baten modura etiketatuta dago jatorrizko zuhaitz-bankuan eta horren parekoa den DU ereduko zuhaitz-bankuan, aldiz, banaturik ageri da eta bere osagaiak *mwe* (*multiword expression*, HAUL) dependentzia-erlazioarekin lotuta.




**4. irudia.** Goian, jatorrizko zuhaitz-bankuko *Ontziak egiteko, batez ere tornuak erabiltzen ziren*. esaldiaren egitura sintaktikoa. Behean, goikoaren parekoa den DUen zuhaitz-bankuko esaldiaren egitura sintaktikoa [13].

### 2.2.2. *Morfologia*

DUen eskeman, hiru mailatan egiten da hitz sintaktiko baten zehaztapena: i) lema: hitzaren eduki semantikoa adierazten du, ii) gramatika-kategoria: hitzarekin lotuta dagoen kategoria lexikal abstraktua adierazten du eta iii) ezaugarri multzoa: hitz-forma bakoitzari dagozkion ezaugarri lexikal eta gramatikalak dira.

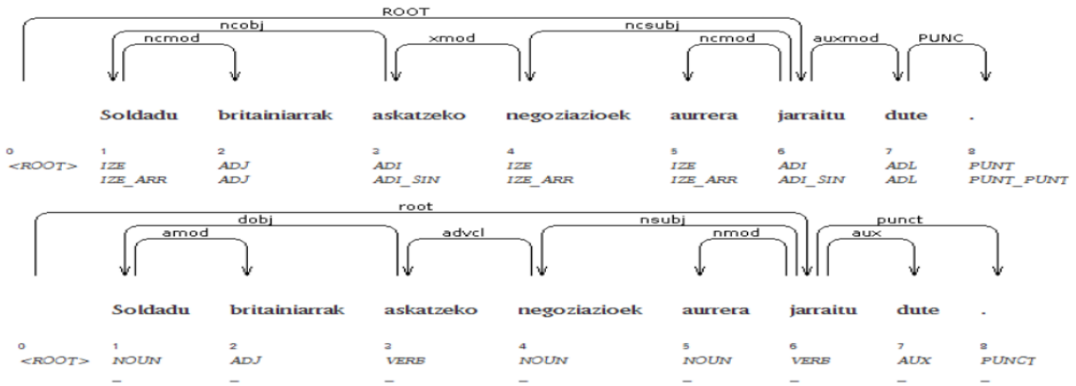
Kategoria-etiketen baliokidetasun-taulan (1. taula) azaltzen da DU ereduan eta jatorrizko zuhaitz-bankuan erabili diren etiketen arteko aldea zein den.

Taularen lehen zutabean ikusten den moduan, DUetako kategoria-etiketa gehienek dute baliokide bat jatorrizkoan, 17tik 10ek. Dena den, egokitzapena ezin izan da zuzenean egin jatorrizko zuhaitz-bankuko beste zenbait etiketak –izenak eta determinatzaileak– baliokide bat baino gehiago dituztelako edota bi etiketa desberdinek baliokide bera dutelako DUetan; kasu horietan guztietan, kategoriaz gain bestelako ezaugarri morfologikoak ere kontuan hartu behar izan dira. Postposizioei (ADP) eta menderagailuei (SCONJ) dagozkien etiketak ere ezaugarri morfologikoetatik inferitu dira.

**1. taula.** DUetan eta jatorrizko zuhaitz-bankuan erabilitako kategoria-etiketak.

| Baliokidetasuna | DU ereduko kategoria-etiketak | Jatorrizko kategoria-etiketak |
|---|---|---|
| 1:1 | ADJ (adjektiboa) | ADJ |
| | ADV (adberbioak) | ADB |
| | AUX (aditz laguntzailea) | ADL |
| | CONJ (juntagailuak) | LOT |
| | DET (determinatzaileak) | DET |
| | INTJ (interjekzioak) | ITJ |
| | PART (partikulak) | PRT |
| | PRON (izenordainak) | IOR |
| | SYM (sinboloak) | SNB |
| | X (bestelakoak) | BST |
| 1:2 | NOUN (izenak) | IZE/LAB (izenak/laburdurak) |
| | NUM (det. zenbatzaileak) | DET/IZE |
| | PROPN (izen bereziak) | IZE/SIG (izen bereziak/siglak) |
| | VERB (aditzak) | ADI/ADT (aditzak/aditz trinkoak) |
| 1:7 | PUNCT (puntuazioa) | PUNT_PUNT/PUNT_KOMA/PUNT_PUNT_KOMA/PUNT_BI_PUNT/PUNT_ESKL/PUNT_GALD/PUNT_HIRU |
| 1:0<br>1:0 | ADP (preposizioak eta postposizioak) | - |
| | SCONJ (menderagailuak) | - |

Kategoria-etiketen egokitzapenaren adibideetako bat da 5. irudian ageri den *Soldadu britainiarrak askatzeko negoziazioek aurrera jarraitu dute.* esaldiaren dependentzia-zuhaitza. Irudi horretako goiko aldean, esaldiaren jatorrizko zuhaitz-bankuko dependentzia-zuhaitza ageri da eta beheko aldean, DUetara egokitutakoarena.




**5. irudia.** Goian, jatorrizko zuhaitz-bankuko *Soldadu britainiarrak askatzeko negoziazioek aurrera jarraitu dute.* esaldiaren dependentzia-zuhaitzeko kategoria- eta dependentzia-etiketak. Behean, goikoaren parekoa den DU zuhaitz-bankukoarena [13].

Maila morfologikoan, HAUL askoren egokitzapena ez da gauzatu, ezin izan ditugulako ziurtasunez banatu osagaiak dagozkien analisiekin. Jatorrizko zuhaitz-

bankuan dauden 1.282 HAULetatik 536 egokitu ditugu. Egokitu gabe utzi dugun HAULetako bat “bien bitartean” da, lehen osagaiaren analisia osatzeko ezinezkoa gertatu zaigulako lehen osagaiaren lema, kategoria, azpikategoria eta ezaugarri morfologikoak berreskuratzea. Ondorioz, HAUL horiek zeuden esaldiak baztertu egin behar izan ditugu.

2.2.3. *Sintaxia*

DUen eskemako etiketatze sintaktikoa esaldietako hitzen arteko dependentzia-erlazioak etiketatzean datza. Esaldi batean, hitz edo token bakoitza beste hitz edo token baten mendekoa da edo esaldiaren erro hipotetikoaren (ROOT) mendekoa. Dependentzia-erlazioen etiketen helburua da hainbat hizkuntzatan erabili diren etiketak modu orokorrean aztertzea eta dependentzia unibertsalen multzoa osatzea; hain zuzen ere, hainbat hizkuntzatako gramatika-erlazio berak modu berean etiketatzeko paralelismoa handitzea bilatzen dute.

DUen ereduan definitu diren 40 dependentzia-etiketak dira baliokidetasunen 2. taularen bigarren zutabean ageri direnak eta hirugarren zutabean, berriz, horien baliokide diren jatorrizko zuhaitz-bankukoak; DUetako etiketaren batek euskaraz baliokiderik ez duela adierazteko – ikurra erabili da.

Sintaxi mailan ere ezin izan da dependentzia-etiketa guztien egokitzapena zuzenean egin, hainbat dependentzia-etiketak baliokide bat baino gehiago dituztelako euskarakoan. Guztietan konplexuena *ncmod* dependentzia-etiketa gertatu da DUetako sei dependentzia-etiketen baliokide delako. Kasu horietan, etiketaz gain, beste ezaugarri batzuk hartu dira kontuan, esaterako mendekoaren gobernatzailea izen edo aditz kategoriako hitza den.

**2. taula.** DUetan eta jatorrizko zuhaitz-bankuan erabilitako dependentzia-etiketak.

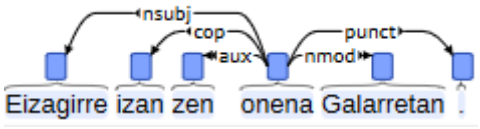
| Baliokidetasuna | DU ereduko dependentzia-etiketak | Jatorrizko zuhaitz-bankuko dependentzia-etiketak |
|---|---|---|
| | **Aditzaren mendekoak eta sintagmaren edo mendeko perpausaren buruak direnak** | |
| 1:1 | nsubj (izen-subjektua) | ncsubj |
| 1:0 | nsubjpass (izen-subjektu pasiboa) | - |
| 1:1 | dobj (objektu zuzena) | ncobj |
| 1:1 | iobj (zeharkako objektua) | nczobj |
| 1:2 | csubj (mendeko perpausa, subjektua) | ccomp_subj/xcomp_subj |
| 1:0 | csubjpass (mendeko perpausa, subjektu pasiboa) | - |
| 1:2 | ccomp (mendeko perpaus osagarria) | ccomp_obj/xcomp_obj |
| 1:0 | xcomp (mendeko perpaus osagarri irekia) | - |
| | **Aditzaren mendekoak eta argumentu ez direnak** | |
| 1:2 | nmod (adizlaguna) | ncmod (adberbioak ez diren adizlagunak) /postos |
| 1:3 | advcl (mendeko perpausa, adizlaguna) | cmod/xmod/ (perpaus adberbialak) /xcomp_zobj |
| 1:2 | advmod (adberbio-modifikatzailea) | ncmod (adberbioak) /gradmod |
| 1:1 | neg (ezezko modifikatzailea) | ncmod |
| | **Perpausaren mendeko bereziak** | |
| 1:1 | vocative (bokatiboa) | itj_out |
| 1:1 | discourse (diskurtsoa) | lot_at |
| 1:0 | expl (espletiboa) | - |
| 1:2 | aux (aditz laguntzailea) | auxmod/galdemod |
| 1:0 | auxpass (aditz laguntzailea pasiboa) | - |
| 1:1 | cop (kopula) | ncpred |
| 1:1 | mark (markatzailea) | menos |
| 1:1 | punct (puntuazioa) | punct |
| | **Izenaren mendekoak** | |
| 1:1 | nummod (zenbakizko modifikatzailea) | detmod (zenbakiak: *3 etxe* modukoak) |
| 1:1 | appos (aposizioa) | aponcmod |
| 1:1 | nmod (izen-modifikatzailea) | ncmod (izenlagunak) |
| 1:2 | acl (adjektibo-perpausa) | cmod/xmod (erlatibozko perpausak) |
| 1:1 | amod (adjektibo modifikatzailea) | ncmod (adjektiboak) |
| 1:2 | det (detrminatzailea) | detmod/ncmod (kantitatea adierazten duten determinatzaile zenbatzaile zehaztugabeak: *asko*…) |
| 1:0 | neg (ezezko modifikatzailea) | - |
| | **Hitz elkartuetako osagaiak eta aztertu ezinak** | |
| 1:1 | compound (hitz elkartua) | ncmod (hitz elkartuak eta zenbakiak: *lau mila* modukoak) |
| 1:1 | name (entitate-izena) | entios |
| 1:1 | mwe (hitz anitzeko unitate) | haos |
| 1:0 | foreign (atzerriko hitzak) | - |
| 1:0 | goeswith (oker banatutako hitzak) | - |
| | **Koordinazioa** | |
| 1:1 | conj (juntagailuak) | lot |
| 1:1 | cc (koordinazio-juntagailuak) | lot |
| 1:1 | punct (puntuazioa) | lot |
| | **Preposizioak eta posesiboak** | |
| 1:0 | case (kasu-markak: preposizioa…) | - |
| | **Zehaztugabeko dependentzia-loturak** | |
| 1:0 | list (zerrenda) | - |
| 1:0 | dislocated (elementu lokatuak) | - |
| 1:1 | parataxis (parataxia) | apocmod |
| 1:0 | remnant (elipsia) | - |
| 1:0 | reparandum | - |
| | **Beste batzuk** | |
| 1:0 | dep (zehaztugabeko dependentzia) | - |
| 1:1 | root | root |

Horretaz gain, zenbait egitura sintaktikoren egokitzapena egin aurretik, esaldi horien analisia aldatu behar izan da DUetara egokitzeko. Jarraian, egitura horietatik

konplexuenak direnak edo egokitzapen-lan handiagoa eskatzen dutenak azalduko ditugu adibideen bitartez.

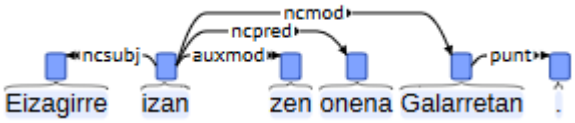
- Aditz kopulatiboak

DUen ereduan, egitura kopulatiboetako gobernatzailea ez da perpauseko aditza, baizik eta osagarria. Horren adibidea da 6. irudiko perpausaren dependentzia-zuhaitza; dependentzia-zuhaitz horretan, *onena* hitza da perpauseko gobernatzailea, eta gainerako hitzak eta puntua bere mendekoak.




**6. irudia.** *Eizagirre izan zen onena Galarretan.* esaldiaren dependentzia-zuhaitza DUetan.

Jatorrizko zuhaitz-bankuko mota horretako egituretan, beraz, perpauseko hitzen arteko dependentzia-erlazioak aldatu behar izan dira, aditza hartzen baitzen gobernatzailetzat (ikus 7. irudia).

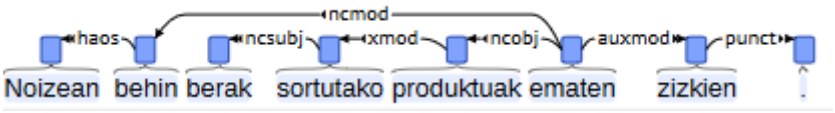



**7. irudia.** *Eizagirre izan zen onena Galarretan.* esaldiaren dependentzia-zuhaitza jatorrizko zuhaitz-bankuan.

- Hitz anitzeko unitateak

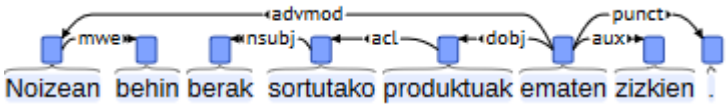
Hitz anitzeko esapideei dagokienez, jatorrizko zuhaitz-bankuan hitz anitzeko unitate lexikalak banatuta zeudenean, bai ziurrak ez zirelako, bai HAULtzat hartu ez zirelako, erabaki zen batetik, HAULaren ezkerreko osagaia eskuineko osagaiaren mendekoa izatea eta *haos* (hitz anitzeko osagaia) dependentzia-etiketaren bitartez lotzea eta, bestetik, HAULaren azken osagaia perpauseko gobernatzailearen mendekoa izatea eta dagokion dependentzia-etiketaren bitartez lotzea. Esaterako, 8. irudiko perpausaren dependentzia-zuhaitzean ikus daitekeen moduan, *noizean behin* hitz anitzeko unitate lexikalaren lehen osagaia –*noizean*– da eskuineko osagaiaren –*behin*– mendekoa, eta eskuineko edo azken osagai hori, berriz, perpauseko *ematen*

aditzaren mendekoa. Kasu honetan, azken osagai hori *ncmod* dependentzia-etiketaren bitartez lotzen zaio perpauseko gobernatzailea den *ematen* aditzari.




**8. irudia.** *Noizean behin berak sortutako produktuak ematen zizkien.* esaldiaren dependentzia-zuhaitza jatorrizko zuhaitz-bankuan.

DUetan, berriz, lehen osagaia da gobernatzailea. Horrela, bada, *noizean behin* hitz unitatean, *noizean* hitza da gobernatzailea eta *behin* hitza bere mendekoa (9. irudia).



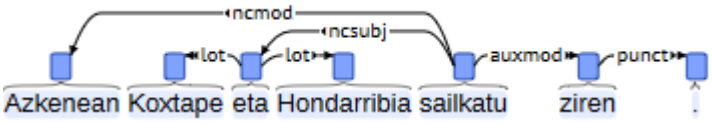

**9. irudia.** *Noizean behin berak sortutako produktuak ematen zizkien.* esaldiaren dependentzia-zuhaitza DUetara egokitu ondoren.

Hitz anitzeko unitate lexikalen antzera bihurtu dira entitate-izenak; horietan ere lehen osagaia hartu da gobernatzailetzat eta bigarrena horren mendekotzat. Nolanahi ere, lehen urrats honetan erabaki dugu maiztasun handieneko eta 2/3 osagaiko hitz anitzeko unitateak egokitzea DUen eredura, hurrengo urratserako utziz konplexuagoak diren unitateak, hau da, maiztasun txikiagokoak eta hiru osagai baino gehiagokoak. Hitz anitzeko unitatetzat hartzen diren hitz elkartuen egokitzapena errazagoa gertatu da.

- Koordinazioa

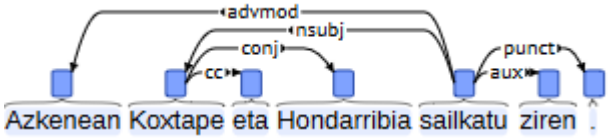
Perpaus edo sintagma koordinatuen analisia ez da modu berean egin dependentzietan oinarritutako zuhaitz-bankuetan. Hala, hizkuntza batzuek elkartzen diren bi elementuetako bat gobernatzailea eta bestea bere mendekoa izatea erabaki dute, eta beste zenbait hizkuntzak, berriz, koordinaturiko bi elementuak biltzen dituen beste adabegi bat txertatzea erabaki dute. Euskararen kasuan erabaki genuen juntadura bideratzen duen juntagailua hartzea adabegitzat eta horren mendeko izatea juntatzen diren beste elementuak. Beraz, 10. irudiko analisian ikusten den moduan, bi sintagma

elkartzen dituen *eta* juntagailua da gobernatzailea eta bere bi mendekoak *Koxtape* eta *Hondarribia*.




**10. irudia.** *Azkenean Koxtape eta Hondarribia sailkatu ziren.* esaldiaren dependentzia-zuhaitza jatorrizko zuhaitz-bankuan.

DUetan, berriz, elkartzen diren elementuetatik lehena hartzen da gobernatzailetzat eta gainerakoak bere mendekotzat (11. irudia). Ondorioz, gure erabakia aldatu behar izan dugu eta hala, kasu honetan *Koxtape* da juntadurako gobernatzailea eta *Hondarribia* bere mendekoa.




**11. irudia.** *Azkenean Koxtape eta Hondarribia sailkatu ziren.* esaldiaren dependentzia-zuhaitza DUetara egokitu ondoren.

## 2.3. Emaitza

Jatorrizko zuhaitz-bankutik besterako egokitzapena bukatu ondoren, jatorrizko zuhaitz-bankuak dituen 150.000 hitzetik 121.443 hitz egokitzea lortu dugu, 8.993 esaldi. Egokitzapen-prozesua automatikoki egin da [14] eta konplexua eta luzea izan da. Kontuan izan behar da egokitzapen-prozesua ez dela izan euskarazko etiketak hartzea eta beraiei dagozkien DUen ereduko etiketetara egokitzea soilik. Prozesu horretan hainbat atal daude: HAULen osagaien analisiak lortu, puntuazio-markak moldatu, koordinazioa moldatu, etiketa bakoitzerako kasuak aztertu eta bere DUetako baliokidea identifikatu, ordena zuzena erabaki... Atal batzuetan, euskarazko etiketen bihurketa zuzena egin bada ere, beste batzuetan zuhaitzaren egitura aldatu behar izan da etiketen bihurketa aplikatu baino lehen.

Egokitzapen-prozesuaren ondotik, bi zuhaitz-bankuen emaitzak konparatu dira Labeled Attachment Score (LAS) neurria baliatuta [12, 13]. Neurri horren bitartez,

hitz kopuru guztietatik gobernatzaileen eta mendekoen arteko dependentzia-erlazio zuzenen portzentajea kalkulatzen da. Jatorrizko zuhaitz-bankuarekin testaren gainean lortutako oinarrizko emaitza % 83koa da eta egokitutakoarekin lortutakoa % 78,50ekoa. Esan daiteke egokitzapenean zerbait galdu den arren, jatorrizko zuhaitz-bankuaren ezaugarri garrantzitsuenei eutsi zaiela.

## 3. ONDORIOAK

Euskarazko jatorrizko zuhaitz-bankua Dependentzia Unibertsalen eredura egokituta, euskara Hizkuntzaren Prozesamenduan kokatzen den nazioarteko proiektu garrantzitsu horren partaide izatea lortu dugu. Egokitzapen-lana egingarria gertatu da zenbait ezaugarri komun dituztelako: biek, DUen ereduak zein jatorrizko zuhaitz-bankuak, jarraitzen dute sintaxiaren hurbilpen lexikalista (erlazioak zatitu gabeko hitz-formen artean gertatzen dira) eta bat datoz eduki-hitzak hartzean izen-sintagmen eta aditz-kateen burutzat.

Horretaz gain, jatorrizko zuhaitz-bankuko kategoriak eta ezaugarri morfosintaktikoak automatikoki egokitzeko prozesua nahiko gardena eta zailtasun handirik gabekoa izan da. Ikusi batera 150.000 hitzetik ia 30.000 hitz gelditu direla egokitu gabe esan badaiteke ere, ez dira zehatz-mehatz 30.000 hitz utzi. Izan ere, irizpide nagusietako bat izan da esaldi bateko hitz bat ezin bada egokitu, esaldi osoa baztertu behar dela. Hori kontuan hartuta, beraz, 6.500 hitz baino gutxiago izan dira egokitu ez direnak, gainontzekoak horien ondorioz baztertu baitira.

Egokitu ez diren hitz horiek lotuta daude dependentzia-erlazioak egokitzerakoan aurkitu ditugun desberdintasunekin eta, ondorioz, egin behar izan ditugun moldaketekin egokitzapena ahalik eta egokiena izan dadin. Horren adibide ditugu Hitz Anitzeko Unitate Lexikalak, besteak beste.

Nahiz eta zuhaitz-bankuaren kalitatea ona izan, badakigu hainbeste mila hitz automatikoki bihurtzen diren prozesuan erroreak egon daitezkeela. Ondorioz, etorkizuneko lanen artean emaitza ondo aztertzeak garrantzia izango du. Halaber, egokitu ez diren HAULak aztertu behar dira zuhaitz-bankuaren tamaina handitzeko. Horretarako, erarik errazena da tratatu gabe geratu diren HAULak modu egokian bihurtzea; horrela ez dira hainbeste esaldi baztertuko.

Bukatzeko, aipatzekoa da zuhaitz-bankuaren erabilgarritasuna. Eredu berera egokitutako zuhaitz-banku horiek erabiliko dira bai analizatzaile sintaktikoak garatzeko, bai analizatzaile horien emaitzen hizkuntzen arteko erkaketa hobetzeko eta bai hizkuntzen tipologiaren araberako egitura sintaktikoen antzekotasuna neurtzeko.

**ESKER ONAK**

Ikerketa-lan hau MECek finantzatutako PROSA-MED (TIN2016-77820-C3-1-R) proiektuaren barruan egin da.

**4. BIBLIOGRAFIA**

[1] NIVRE, J. 2014. «Universal Dependency Parsing», (Invited talk). *First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages* (SPMRL-SANCL).

[2] PETROV, S., DAS, D. eta MCDONALD, R. 2012. «A Universal Part-of-Speech Tagset». *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), 2089 – 2096.

[3] ROSA, R., MAŠEK, J., MAREČEK, D.. POPEL, M., ZEMAN, D. eta ŽABOKRTSKÝ, Z. 2014. «HamleDT 2.0: Thirty Dependency Treebanks Stanfordized». *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), 2334 – 2341.

[4] ZEMAN, D., DUŠEK, O., MAREČEK, D., POPEL, M., RAMASAMY, L., ŠTĚPÁNEK, J., ŽABOKRTSKÝ, Z. eta HAJIČ, J. 2014. «HamleDT: Harmonized multi-language dependency treebank». *Language Resources and Evaluation*, **48**, 601 – 637.

[5] DE MARNEFFE, M.C. eta MANNING, C. D. 2008. «The Stanford typed dependencies representation». *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*, 1 – 8.

[6] DE MARNEFFE, M.C., DOZAT, T., SILVEIRA, N., HAVERINEN, K., GINTER, F., NIVRE, J. eta MANNING, C. D. 2014. «Universal Stanford dependencies: A cross-linguistic typology». *Proceedings of 9th International Conference on Language Resources and Evaluation* (LREC 2014), **14**, 4585 – 4592.

[7] NIVRE, J., AGIĆ, Ž., AHRENBERG, L. et al. 2017. *Universal Dependencies 2.1*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL). Faculty of Mathematics and Physics, Charles University. Praga. http://hdl.handle.net/11234/1-2515 (2018-05-10).

[8] ARANZABE, M.J. 2008. *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala*. Doktorego-tesia. Euskal Hizkuntza eta Komunikazioa Saila, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU).

[9] ADURIZ, I., ARANZABE, M.J., ARRIOLA, J.M., ATUTXA, A., DÍAZ DE ILARRAZA, A., EZEIZA, N., GOJENOLA, K., ORONOZ, M., SOROA, A. eta URIZAR, R. 2006. «Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing». In Andrew Wilson, Paul Rayson and Dawn Archer (arg.), *Corpus Linguistics Around the World*, **56**, 1 – 15

[10] BENGOETXEA, K. eta GOJENOLA, K. 2016. «Euskarako analizatzaile sintaktiko-estatistikoa hobetzeko teknikak». *Ekaia*, **ale berezia**, 19 – 45.

[11] URIZAR, R. 2012. *Euskal lokuzioen tratamendu konputazionala*. Doktorego-tesia. Euskal Hizkuntza eta Komunikazioa Saila, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU).

[12] NIVRE, J. eta FANG CH.-T. 2017. «Universal Dependency Evaluation». *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, 86–95.

[13] GOENAGA, I. 2017. ASKHI: *Analisi sintaktiko konputazional hibridoa paradigma desberdinen konbinazioan oinarrituta*. Doktorego-tesia. Lengoaia eta Sistema Informatikoak Saila, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU).

[14] ARANZABE, M.J., ATUTXA, A., BENGOETXEA, K., DÍAZ DE ILARRAZA, A., GOENAGA, I., GOJENOLA, K. eta URIA, L. 2015. «Automatic Conversion of the Basque Dependency Treebank to Universal Dependencies». In Markus Dickinsons, Erhard Hinrichs, Agnieszka Patejuk, Adam Przepiórkowski (arg.), *Proceedings of the Fourteenth International Workshop on Treebanks an Linguistic Theories (TLT14)*, 233–241.