Euskarazko Hitz Anitzeko Unitate Lexikalen tratamendu konputazionala

Ruben Urizar, Iñaki Alegria, Juan Carlos Odriozola, Nerea Ezeiza*

Laburpena

Multi-word Lexical Units (MWLU) are of great importance in language in general, and in Natural Language Processing in particular, since they are not governed by the free rules of the system. In this article, we give an overview of the different types of phraseological units, explaining briefly each one's features. Being our priority to process idioms automatically in Basque texts, we concisely analyze several approaches for the inflectional description of MWLUs, and then, we explain the system we have developed for Basque: (i) a general representation for describing MWLUs in the lexical database for Basque (EDBL), (ii) HABIL, a tool capable of detecting and analyzing them based on the features described in the database, and (iii) a constraint grammar for disambiguating ambiguous MWLUs.

1. Sarrera

> La formación, el funcionamiento y el desarrollo del lenguaje están determinados no sólo por las reglas libres del sistema, sino también por todo tipo de estructuras prefabricadas de las que se sirven los hablantes en sus producciones lingüísticas. (Corpas, 1996).

Gloria Corpasen hitzok argi uzten dute sintaxi asketik kanpo finkaturik dauden egiturek hizkuntzan duten garrantzia. Izan ere, Jackendoffen (1997) estimazioen arabera, hiztunen lexikoan, hitz bakunen besteko munta dute hitz anitzeko unitateek. Beste autore batzuek are proportzio handiagoak ematen dituzte.

Euskararen prozesamendu konputazionalean, IXA taldeak garatua du mendekotasun-egituretan oinarrituriko *parser* edo analizatzaile sintaktiko sendoa (Aduriz *et al.*, 2004). Analizatzaileak mailaka gauzatzen du analisia, kate-maila bakoitzean hizkuntza-ezagutza desberdina baliatuz, hasi tokenizaziotik eta mendekotasun sintaktikoen erlazioak ezarri arte. Hitz Anitzeko Unitate Lexikalen (HAUL) tratamendua berebiziko garrantzia du prozesu horretan.

* Halaber, biziki eskertu nahi dugu Kike Fernández, HABILen inplementazioan egindako lan eskergagatik

2. Unitate fraseologikoak eta HAULak

HAULaren definizioa eta hizkuntzaren prozesamenduan tratatu beharreko unitate motak nabarmen aldatzen dira helburuen arabera eta burutu nahi den prozesuaren sakontasunaren arabera (Copestake *et al.*, 2002).

Txitean-pitean darabilgun *hitz* kontzeptua bera ez da argia, inondik ere. Bestelako hizkuntza-auziak alde batera utzita, hizkuntzaren prozesamenduan, eta testu mailan, unitate *grafikoa*ri egiten dio erreferentzia. Alegia, *bereizle*en (*separator*) arteko karaktere-katea da hitza (Savary, 2008). Bereizleak karaktere ez-alfabetikoak izan ohi dira —nagusiki zuriuneak eta puntuazio-markak— baina aldatu egiten dira hizkuntzatik hizkuntzara edo ikerlanen arabera. Japonieraz edo Thailandieraz, esate baterako, zuriunerik gabe idazten da. Halaber, zenbait ikerlanetan (Silberztein, 1993) apostrofoa edo marratxoa hitz-bereizletzat hartzen dira eta, hala, ingelesezko *don't* edo frantsesezko *ajourd'hui*, esate baterako, bi hitzez osatuak lirateke. Beste ikerlan batzuetan, berriz, apostrofoa edo marratxoa hitz barruko osagaia izatea onartzen da, eta, hortaz *don't* eta *aujourd'hui* hitz bakartzat hartzen dira (ikus "tokenizazioa" 4. atalean).

Bestalde, euskara bezalako hizkuntza eranskarietan unitate tipografiko bakarra osatzen duten hainbat hitz HAULak lirateke flexiorik gabeko hizkuntzetan (*aurrerantzean* 'en adelante', *baraurik* 'à jeun', *ziurrenik* 'most probably'). Beraz, euskaraz horiek hitz bakarra balira bezala tratatzen dira, datu-base lexikalean hitz bakun gisa sarrera emanaz.

Corpasek (1996) hiru multzo edo *esfera* nagusitan banatzen ditu unitate fraseologikoak, alegia, sistemaren erregela askeetatik at dauden egiturak: kolokazioak, lokuzioak eta enuntziatu fraseologikoak.

Konbinazio libreek, kolokazioek, eta lokuzioek, ostera, continuum bat osatzen dute, eta sarri askotan, euren arteko mugak ez dira batere argiak izaten. Hala ere, unitate fraseologiko horien definizio eta ezaugarrien berri ematen saiatuko gara ondoren, batez ere Corpasen (1996) irizpideei jarraiki.

Kolokazioak

Zale amorratua, arrain sarda, izerdia bota, pozez zoratu edo *politikoki zuzena* bezalako konbinazioei deritze kolokazioak. Erlazio sintaktikoren bat duten bi unitate

lexikok osatzen dute (izen-izenondo, izen-izen, izen-aditz, adberbio-aditz, adberbio-adjektibo...). Erabilerak ezarritako *konbinazio-murriztapenak* dituzte: *oinarriak* (adib. *zale*), semantikoki autonomoa izanik, *kokakide*aren (*amorratu*) hautapena determinatzen du, eta gainera, honen adiera berezi bat hautatzen du, sarritan abstraktua edo figuratiboa dena ('grinatua'). Bestalde, morfosintaktikoki konposizionalak eta aldagarriak dira, eta, hau dela eta, zenbait aldaki formal onartzen dituzte, hala nola, kokakidea ordezkatzea (*zale amorratu/sutsu/porrokatu*), modifikatzailea gehitzea (*izerdi hotza egin*), nominalizazioa (*zuzentasun politikoa*), erlatibizazioa (*bota dudan izerdia*) etab.

Lokuzioak

Idiomatikotasuna edo/eta zurruntasun formala ezaugarri duten egitura finkoei deritze lokuzio. Batetik, esapide hauetako asko semantikoki ez-konposizionalak dira, hau da, horien interpretazioa nekez egin daiteke hura osatzen duten osagaien esanahietatik abiatuta (*hanka sartu, begitan hartu, zorri piztua, batez beste, hala ere, behinik behin*). Beste batzuetan, ostera, predikatu konplexu batzuetan kasu (*lan egin, min hartu, berri eman*), lokuzioaren esanahia osagaien esanahiak konbinatuz lor daiteke hein handi batean. Euren sintaxi idiosinkratikoak bihurtzen ditu lokuzio. Izan ere, Koikeren (2001) hitzetan, "*en la fijación estructural se basan todos los comportamientos de las locuciones*". Zabalak (2004) honelako zenbait egituretan gertatzen diren berezitasun sintaktiko batzuen berri ematen du, argi utziz lokuzio guztiek ez dutela zertan ezaugarri guztiok eta neurri bertsuetan bete behar.

Esate baterako, objektu zuzen arruntak euskaraz beti determinatzaileren bat eskatzen badu ere (1), *lan egin* bezalako hainbat aditzetan, ustez objektua denak ez du mugatzailerik hartzen (2).

(1) *Arroz jan dut

(2) Jonek goizean lan egin du

Maila indibidualeko predikatuak ere mugatuta joan ohi dira (Zabala, 1993); horregatik (3) adibidea ez da gramatikala. Ostera, *zilegi izan* predikatu konplexua mugatzailerik gabe doa beti (4).

(3) *Harrigarri da zuk hori esatea

(4) Zilegi da hori egitea

Beste kasu batzuetan, konposatuaren aditzak azpikategorizazio-ezaugarri berria bereganatzen du, bere kabuz doanean ez daukana. Berbarako, *ekarri* aditzak normalean ez du mendeko perpaus konpletiborik azpikategorizatzen; *gogora ekarri*-k, aldiz, bai.

(5) *Gogora ekarri* zidan lana egin behar nu*ela*

Halaber, ez da ohikoa *egin* aditz iragankorrak adizlaguna hartzea objektu zuzenaren ordez, *hegaz egin* edo *eztulka egin* bezalako hainbat predikatu konplexutan gertatzen den moduan.

Bestalde, sekuentzia hauetako askotan agertzen diren elementu batzuk ez dira azaltzen beste inongo testuingurutan, eta sarritan zein kategoriatakoak diren zehaztea ere zaila gertatzen da. Adibidez, *noizik* hitza bakarrik agertzen da *noizik behin, noizik behinean, noizik noizera* eta *noizik behinka* esamoldeetan; *behinik* hitza *behinik behin* HAULean azaltzen da soilik; zer kategoriatakoak dira *zirt*, *zart*, *trikun-trakun*, *laprast* edo *fio*, ia esklusiboki *zirt edo zart egin, trikun-trakun egin, laprast egin* eta *fio izan* aditzetan agertzen direnak hurrenez hurren?

Beste zenbaitetan, fosildurik dagoelako-edo, ezinezkoa gertatzen da hitz-konbinazioaren interpretazioa (semantikoa edo sintaktikoa) egitea osagaiek egun duten kategoria (edo esanahia) kontuan izanik. Esate baterako, *beldur —gose* edo *egarri* ez bezala— izena da soilik; *beldur naiz* (*beldur izan*) HAULa ulertzeko, ostera, izenondo (edo aditzondo) interpretazioa behar luke (6). Berdintsu gertatzen da *mintzo izan*ekin (7). (8) esaldia sintaktikoki analizatu ahal izateko, berriz, elipsi bat-edo gertatu dela pentsatu behar genuke: *egarriak "jota" nago*.

(6) Oso beldur ginen

(7) Gipuzkeraz mintzo da

(8) Emadazu edaten ur pittin bat, egarriak nago eta

Enuntziatu fraseologikoak

Azkenik, enuntziatu fraseologikoak esaldi osoak dira. Corpasek (1996) hiru multzo nagusitan banatzen ditu hirugarren esferako unitateok: paremiak (*balizko olak burdinarik ez, urak dakarrena urak darama*), formula diskurtsiboak (*Zer moduz?, Egun on*) eta formula psikosozialak (*Sentitzen dut, Bai horixe, Zer egingo diogu?, Hor konpon, Alde hemendik, Bizi al gara?*).

Gure ikerlanaren fase honetan, ez ditugu enuntziatu fraseologikoak aintzat hartu, ezta kolokazioak ere, lokuzioak lantzeak lehentasuna zuela iritzi baitiogu[1]

Datu-base lexikalean zein hitz anitzeko unitate landu erabakitzerakoan, lexikografoen ezagutzaz baliatu gara, eta bi iturritatik edan dugu, batez ere, orain arteko lanean. Batetik, *XX. Mendeko Euskararen Corpus Estatistikoa*[2] erabili dugu, eta bertan lematizaturiko HAULen artean maiztasun handienez ageri direnak hautatu ditugu. Bestetik, Euskaltzaindiaren *Hiztegi Batu*ko (Euskaltzaindia, 2002) hitz anitzeko sarrerak hartu ditugu. Guztira 2300 bat HAUL landu dira.

3. Hitz anitzeko unitateen prozesamendurako hurbilpen lexikoak

Savaryk (2008) HAULen flexioak adierazteko zenbait hurbilpen lexikoren azterketa kontrastiboa egiten du. Azterketa horretan, lau multzo nagusitan banatzen ditu ikerlanak: DELA hiztegiak, egoera finituak kaskadan (*cascaded finite-state*), baterakuntza-gramatikak eta datu-base erlazionalak.

3.1. DELA hiztegiak

Parisko DELA eskolaren hiztegi elektronikoetan (Courtois eta Silberztein, 1990) hitz bakunak eta izen-lokuzioak sistematikoki zerrendatzen dira eta horien flexioaren deskribapena egiten da banan-banan. Alabaina, hitz elkartu jarraituak bakarrik deskribatu dituzte.

Flexio-kode baten bidez adierazten dute osagaien flexio-paradigma guztia eta osagaien arteko genero- eta numero-komunztadura (*toile d'araignée, toiles d'araignée, toiles d'araignées*).

---

[1] Beste modulu batzuetan, datak eta zenbakizko esapideak ere tratatzen dira. Horiek multzo irekia izanik, ez daude datu-basean sarturik. Hala ere, osagaiak beti elkarren ondoan joan ohi dira, egitura finkoak dituzte, eta lexikoi itxia darabilte. Beraz, erraza da horiek identifikatzea egoera finituko transduktore sinple batekin (Ezeiza, 2002).

[2] `www.euskaracorpusa.net`

3.2. Egoera finituak kaskadan

Hurbilpen honetako tresnarik ezagunenak ditugu *xfst*, adierazpen erregularren konpilatzailea, eta *lexc*, egoera finituko lexiko-konpilatzailea (Beesley eta Karttunen, 2003). Bi hauek lan-tresna paregabea eskaini dute HAULak prozesatzeko hainbat ikerlanetan. Horien artean azpimarratzekoak dira Karttunen *et al*-en frantsesezko hitz elkartuen kasu-ikerketa (Karttunen *et al.*, 1992; Karttunen, 1993), IDAREX (Breidt *et al.*, 1996) edo turkierarako HAULen prozesadorea (Oflazer, 1996).

Adibidez, IDAREXek, alemanezko aditz-lokuzioen deskripzioa egiten du, batez ere. Adierazpen erregularren bidez adierazten ditu forma flexionatu posibleak. Osagaien flexio-murriztapenak ez ezik, hautazko osagaiak, bikoizketak edo elementuen tartekatzeak ere deskribatzen ditu, bai eta transformazio sintaktikoak ere. Elementuren bat tartekatzen denean, zehaztu egin behar da elementu hori zein den (aditzondoa, izenordaina...).

3.3. Baterakuntza-gramatikak

Zenbait ikerlanek baterakuntzan oinarrituriko gramatikak baliatu dituzte HAULen deskripziorako, osagaien arteko komunztadura ez ezik konputazionalki konplexuagoak diren bestelako dependentziak adierazteko ere egokia delako. Horien artean, *Lingo* proiektua (Sag *et al.*, 2002; Copestake *et al.*, 2002; Villavicencio *et al.*, 2004) eta FASTR (Jacquemin, 2001) dira aipagarrienak.

Ingelesezko hitz anitzeko esapideen deskripzio semantiko eta sintaktikorako proiektu handia da *Lingo*. Sintaxia deskribatzeko gramatikek ezaugarri-egiturak erabiltzen dituzte, murriztapenetan oinarrituriko HPSG formalismo baten baitan. Semantikarako, ostera, *Minimal Recursion Semantics* izenekoa baliatzen dute.

3.4. Datu-base erlazionalak

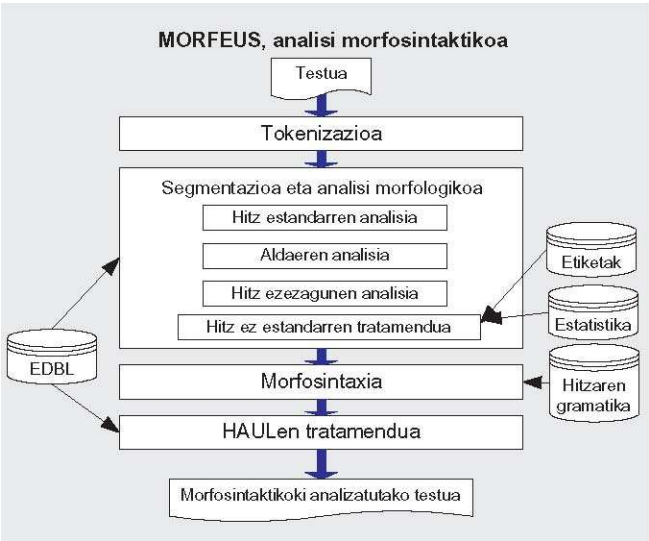
Hitz bakunak ez ezik unitate konplexuak ere deskriba daitezke datu-base erlazional batean, noski, areago unitate konplexuaren osagaiak datu-baseko hitz bakunak izanik.

Euskarazko hitz anitzeko unitateen prozesamendurako garatu dugun modeloan, HAULen deskripzioa datu-base erlazional batean gauzatzen da. Hala ere, datu-

basean eginiko deskribapenetatik adierazpen erregularrak lortzen dira, gero HABIL programak (ikus 5.2. atala) interpretatzen dituena. Ondorengo ataletan azalduko ditugu deskripzio horren nondik norakoak bai eta unitate konplexu horien tratamenduan esku hartzen duten prozesu guztiak.

4. Hitz bakunen tratamendua

Sarreran aipatu dugun moduan, euskarazko analizatzaile sintaktikoak mailaka burutzen du testuen analisia, modulu bakoitzak aurrekoaren emaitzetan oinarritzen baita. Hitz anitzeko unitateen prozesamendua, 1. irudian ikus daitekeenez, hitz bakunen analisien ostean —eta beraz haietan oinarrituta— egiten da.



1. irudia: MORFEUSen eskema orokorra

MORFEUS (Aduriz *et al.*, 1998) analizatzaile morfosintaktikoarekin hasten da analisi-prozesua. MORFEUSen helburua testuko *hitz* orori analisi posible guztiak esleitzea da, bakoitzari, dagokion *lema* eta kategoria gramatikalaz gain, informazio morfologiko egokia esleituz. Horretarako, Ezeizak (2002) azaltzen duenez, 4 pauso nagusi ematen dira: tokenizazioa, segmentazioa eta analisi morfologikoa, morfosintaxia eta HAULen prozesamendua. Atal honetan, bada, hitz bakunen prozesaketan esku hartzen duten hiru moduluak azalduko ditugu, labur bada ere.

1. **Tokenizazioa.**

Edozein testu analizatu nahi bada, egin beharreko lehen pausoa tokenizazioa da, alegia, *token* edo unitateak bereiztea. Jakina da puntuazio-markak, esate baterako,

hitzaren hondarrean itsatsirik ageri direla. Bada, hitza eta puntuazio-markok banatzen ditu tokenizazioak. Baina, hitzak eta puntuazio-markak ez ezik, zenbakiak —arabiarrak zein erromatarrak—, laburdurak, siglak eta abar ere identifikatzen ditu testuan, deklinaturik zein deklinatu gabe egon. Halaber, tokenei zenbait informazio tipografiko gehitzen die etiketen bidez, hala nola, letra larriz hasten den, hitz osoa maiuskulaz idatzita ote dagoen, sigla edo zenbaki deklinatua ote den etab.

Beraz, tokenizazioa urrats garrantzitsua da, izan ere, ondorengo prozesu guztiak unitate horietan oinarriturik daude.

Gainera, ez da kontu hutsala. Tokenak banantzeko erabiltzen diren bereizleak (*separators*) aldatu egiten dira hizkuntzatik hizkuntzara edo ikerlanen arabera. Adibidez, zenbait ikerlanetan (Silberztein, 1993) karaktere ez-alfabetiko guztiak hartzen dira hitz-bereizletzat.

Euskararen kasuan, MORFEUSek ez du marratxoa bereizletzat hartzen. Marratxoaren tratamendua bi modutara gauzatzen da. Zenbait marradun hitzek, lexikalizaturik daudela iritzita, sarrera dute Euskararen Datu Base Lexikalean (EDBL), marratxo eta guzti (*botoi-zulo*), eta beraz, hitz bakunen tratamendu bera dute. Marratxoak, baina, beste erabilera batzuk ere baditu euskaraz, hala nola, izen berezietan oinarri lexikoa eta morfema dependentea bereizteko (*Shakespeare-k*). Kasu horiek —eta konbinazio askez sorturiko hitz elkartuak (*mahai-hanka*)— analizatu ahal izateko marratxoak sarrera du EDBLn "elementu lexiko" gisa, eta hartara MORFEUS, morfotaktika egokiari esker, gauza da horiek ezagutzeko (ikus hurrengo puntua: segmentazioa).

2. **Segmentazioa eta analisi morfologikoa.**

Testu-hitz bakoitza esanahia duten osagai txikienetan (lexema eta morfemetan) zatitzea du helburu segmentazioak, token bakoitzaren interpretazio (analisi) posible guztiak emanaz (Alegria, 1995; Urkia, 1997; Ezeiza, 2002). Halaber, osagai bakoitzari dagokion informazio morfologikoa ere gehitzen zaio, besteak beste, kategoria eta azpikategoria, deklinabide-kasua edo postposizio-atzizkiaren kasua, numero eta mugatasuna edo, aditzen kasuan, modu/denbora, aspektua, pertsona... Hizkuntza-informazio hau guztia EDBLetik jasotzen du segmentatzaileak. Datu-base aberats honek laurogei bat mila sarrera ditu, hiru multzo nagusitan banatuak: hiztegi-sarrerak, adizkiak eta morfema dependenteak (deklinabide-atzizkiak, postposizio-

atzizkiak, erlazio-morfemak, atzizki eta aurrizki lexikalak, aspektu-morfemak, elipsia, elkarketa-marra). Horietako sarrera bakoitzak bere informazio morfologikoa du datu-basean, eta, horrekin batera, bere *morfotaktika*ri buruzko murriztapena, hau da, erantsi dakizkion morfema guztien multzoa eta horien hurrenkera-murriztapenak adierazten dituen *jarraitze-klasea*.

3. **Morfosintaxia.**

Testu-hitz bakoitzaren lexemei eta morfemei buruz segmentazioan lorturiko informazio morfosintaktikoa *biltzea* eta *optimizatzea* da analizatzaile morfosintaktikoaren helburua (Aduriz, 2000; Arriola *et al.*, 2005). Analisi morfosintaktikoa gauzatzen duen gramatika testuingururik gabekoa da, hitzaren barruko egitura deskribatzen baitu (Gojenola, 2000; Aduriz *et al.*, 2000). Gramatikak, lexema eta morfemetatik lortutako informazioa konbinatuz, ezaugarri-egitura bana sortzen du testu-hitzaren interpretazio bakoitzeko. Ezaugarri-egituren arteko bateragarritasuna egiaztatzeko, *baterakuntza* erabiltzen du. Programari dagokionez, PATR formalismoaren inplementazio bat aukeratu da (Shieber, 1986), bere sinpletasun eta malgutasunagatik.

5. Hitz Anitzeko Unitate Lexikalen tratamendua

Arestian azalduriko hitz mailako analisi morfosintaktikoan oinarriturik burutzen da HAULen prozesamendua. Bada, prozesamendu horretarako honakoak garatu behar izan ditugu:

- HAUL horiek guztiak EDBLn deskribatzeko adierazpide orokor bat.
- Datu-basean eginiko deskribapenetan oinarrituta, HAULak (edo HAUL hautagaiak) detektatu eta analizatzen dituen HABIL tresna, HAULen interpretazio guztien analisi morfosintaktikoak sortu eta katean integratuko dituena.
- Murriztapen-gramatika bat, HABILek hautagai gisa aurkezten dituen HAULak desanbiguatuko dituena, alegia, ebatziko duena zein testuingurutan diren egiaz hitz anitzeko unitate, eta zeintzuetan ez.

Ondorengo ataletan azalduko dugu pauso horietako bakoitza.

5.1. HAULak deskribatzeko adierazpidea

HAUL bat EDBLn gehitzean, honako informazio linguistikoa zehazten da datu-basearen taula ezberdinetan:

- Sarrera
- HAULaren kategoria eta azpikategoria
- Barne-osaketa: taula honetan adierazten da zeintzuk diren HAULaren osagaiak, eta zein osagaik ematen dion informazio morfologikoa unitate osoari
- Gauzatze-eskemak: HAULa, testuetan, zein patroien arabera gauzatu daitekeen adierazten da taula honetan

5.1.1. Sarrera. Kategoria eta azpikategoria

Taula hauetan, HAULaren lema kanonikoa eta unitate osoari dagozkion kategoria eta azpikategoria zehazten dira, batez ere.

Hiztegi-sarrera bakunen kategoria berberak izan ditzakete HAUL osoek ere (izen arrunta, adjektiboa, adberbioa, aditza, izenordaina, juntagailua, lokailua, interjekzioa...)[3].

5.1.2. Barne-osaketa

Barne-osaketaren taulan, HAULaren osagaiak zehazten dira. Horretarako osagai bakoitza dagokion hiztegi-sarrera bakunarekin lotzen da. Horrela, flexioa har dezakeen osagaiak lema horren arabera hartuko duela adierazten dugu. Halaber, HAUL osoari zein osagaik ematen dion informazio morfologikoa ere zehazten da. Esate baterako, *aditzera eman* HAULaren lehen osagaia *aditu* sarrera bakunarekin lotua dago eta bigarren osagaia, ostera, *eman* sarrerarekin. Era berean, HAUL osoari informazio morfologikoa emango dion osagaia *eman* dela ere adierazi dugu.

Gerta liteke, baina, HAUL baten forma kanonikoa eta osagaietako batena beti bat ez etortzea. Esate baterako, *maite izan* aditzaren bigarren osagaiak *izan* aditz nagusiaren formez gain (*izan*, *izango*, *izaten*, *izateko*...) (9), *ukan* aditzari

[3] Merezi du aipatzea, baina, hitz anitzekoen osagai gisa bakarrik ager daitezkeen hitzentzat HAOS (hitz anitzekoen osagaia) kategoria sortu genuela, hain zuzen, esana dugun bezala, beti ez baita posible HAUL osagaien kategoria zein den zehaztea (*zirt, zart, trikun-trakun, behinik, fio*).

dagozkionak ere har ditzake, aditz nagusi moduan (*ukan*, *ukanen*...) (10) zein aditz trinko bezala (*zaitu*, *du*, *gintuen*...) (11).

(9) *Maite izatea* hobea da, ikaragarria da, ordea. (*Fantasiazko ipuinak*, G. de Maupas sant / Josu Zabaleta) [4]

(10) Loreak beti *maite ukan* ditut. (*Bihotzeko mina*, Janbattitt Dirassar)

(11) *Maite zaitu* haatik zentzuduna bazina bezala. (*Harreman arriskutsuak*, Ch. de Laclos / Jon Muñoz)

5.1.3. Gauzatze-eskema

2. atalean azaldu dugun moduan, lokuzio askoren ezaugarrietako bat egitura-zurruntasuna bada ere, lokuzio batetik bestera zurruntasun maila asko aldatzen da. Adibidez, lokailu-lokuzioak (*behinik behin, hain zuzen ere, batik bat*) edo adberbio-lokuzio askok (*bide batez, hitzetik hortzera, gehienez ere*) erabateko zurruntasun morfosintaktikoa dute. Aldiz, beste muturrean, *maite izan* moduko aditz-lokuzioek aldagarritasun morfologiko eta sintaktiko handia dute.

(12) [...] ez *duzula* lehendakariaren emaztea gutxiago *maite* (*Harreman arriskutsuak*, Choderlos de Laclos / Jon Muñoz)

(13) [...] Jaungoikoak *maite* ez *dituelako* (*Hainbat idazlan*, San Agustin / Imanol Unzurrunzaga)

Bada, gauzatze-eskemen taulan, patroien bidez deskribatzen da HAULa zein forma desberdinetan gauzatu daitekeen testuetan. Patroi horiek zehazteko hurrenkera, flexio-murriztapenak eta ziurtasuna eremuak baliatzen ditugu.

Hurrenkera

Taula honetan, osagaien hurrenkera zehazten da; patroi bakoitzeko bat. Horretarako, lehen osagaiari 1 zenbakia egokitzen zaio, bigarrenari 2a, eta abar. Osagaien artean HAULekoa ez den hitzik ager badaiteke, ikur bereziak tartekatzen dira zenbakien artean:

\+ (osagai bat edo gehiago)

[4] HAULen erabilerari buruzko argibideak ematean, adibide errealekin hornitzen saiatu gara. Adibideok *Ereduzko Prosa Gaur* corpusetik hartuak dira gehienbat, eta horrelakoetan, liburuaren izenburuarekin batera, egilea eta, hala dagokionean, itzultzailea ere zehaztu ditugu adibideen ostean. Gainerako kasuak Internetetik atereak izanik, erreferentziarik gabe eman ditugu.

* (zero osagai edo gehiago)

? (gehienez osagai bat)

Adibidez, 12 ordenan, HAULaren bigarren osagaia lehenengoaren segidan agertuko da, tartean beste hitzik ez duela.

(14) Badaezpada ere, ez nion *aditzer*a *eman*, behar baino lehenago ikara ez zedin. (*Rock'n'roll*, Aingeru Epaltza).

2*1 ordenan, ostera, bigarren osagaia lehenengoaren aurrean azaltzeaz gain, tartean hitz bat edo gehiago ager daitezkeela esan nahi du.

(15) Idazkerak aztoramendu handia *ematen* du *aditzera* (*Kontakizunak*, E.A. Poe / Koro Navarro)

(16) Besterik gabe *emango* dut, hortaz, *aditzera* ikasle harekin izan nuen harremana (*Ez balego beste mundurik*, Karlos Linazasoro)

Flexio-murriztapenak

Behin osagaien hurrenkera zehaztuta, HAULaren osagaiek zein flexio har ditzaketen zehazten da taula honetan. Osagaia dagoen-dagoenean azal daiteke beti, edo, beste muturrean, edozein flexio onar dezake. Adibidez, *aditzera eman* sarreraren lehen osagaia beti forma berean agertuko da testuetan (*aditzera*), bigarrenak, ostera, edozein flexio hartu ahal izango du (*eman*, *ematen*, *emateko*, *emanaz*, *emaiozu*...).

Berriz, osagaiak flexio jakin batzuk besterik onartzen ez dituenean, halaxe adierazten dugu. Adibidez, *maite izan* aditzaren lehen osagaia absolutibo mugagabean ez ezik (17), graduaturik ere ager daiteke, gradu konparatiboan (18), superlatiboan (19) zein gehiegizkoan (20).

(17) Eromeneraino *maite* izan nuen! (*Fantasiazk*o *ipuinak*, G. de Maupassant / Josu Zabaleta)

(18) [...] nik gero eta *maiteago* ditut boz isilak eta kolore apalak (*Lur bat haratago*, Joan Mari Irigoien)

(19) Ama: nor duzu *maiteen* ni ala Asier? (*Traizioak*, Iban Zaldua)

(20) Ondikotz! neska xarmant bat *maiteegi* ukan zuen [...] (*Gauaz park*e *batean*, Jon Mirande)

Halaber, *Euskal Herri* izen bereziaren azken osagaiak letra larriz eta singularrean joan behar du beti (21). Hala ez balitz, *Euskal Herri* HAULa barik, beste zerbait litzateke (22, 23).

(21) Zalantzarik gabe, *Euskal Herriaren* zatiketa administratiboarekin dauka horrek zerikusia (*Euskara batua*, Koldo Zuazo)

(22) [...] eta *euskal herri* xehearen defentsa du helburu eta ardatz (*Putzu*, Txilardegi)

(23) [...] egia agertu eta argi egin baitzait gure *euskal herrietako* besta gozagarrietan (*Trapuan pupua*, Patziku Perurena)

Flexio-murriztapenak adierazteko, analisietan ageri den edozein informazio erabil daiteke, berez: deklinabide- edo postposizio-atzizkien kasuak, numero eta mugatasuna, aditz mota, ortografia (maiuskula)...

Hauxe litzateke, bada, *maite izan* lokuzioaren lehen osagaiak (*maite*) duen flexio-murriztapena:

```
(KAS=ABS ETA MUG=MG) EDO (MAI=* + (KAS=ABS ETA MUG=MG))
```

Bertan adierazten da osagaiak har dezakeela bai absolutibo mugagabea (*maite*: `KAS=ABS ETA MUG=MG`) bai graduatzaile bat gradu-maila edozein duela (`MAI=*`), beti ere atzetik absolutibo mugagabea duelarik (`KAS=ABS ETA MUG=MG`), hau da, *maiteago*, *maiteegi* edo *maiteen* (18, 20, 19), baina ez, esate baterako, *maiteagoak* edo *maiteenak* (24, 25). Bi adibide hauek, hortaz, ez dira *maite izan* aditz-lokuzioaren agerpenak.

(24) Geroz eta *maiteagoak* dira Miguel bezalako pertsonei esker

(25) Elkarteko aretoan zintzilikatu dituzten argazkiak egilearen irudi *maiteenak* ziren

Ziurtasuna

Azkenik HAUL bakoitzeko patroia *ziurra* ala *anbiguoa* den zehaztuko da. HAULaren osagaiak, patroian zehazturiko hurrenkera eta flexio-murriztapenekin, testu batean agertuz gero ezinbestean hitz anitzeko unitatea izango bada, patroi hori ziurra dela esango dugu.

Esate baterako, *hala eta guztiz ere* hitz anitzeko lokailua ziurra dela esaten dugu zeren eta, *hala, eta, guztiz* eta *ere* osagaiek testu batean ordena eta forma horiexetan agertzen diren guzti-guztietan, hitz anitzeko lokailuaren interpretazioa izango baitute.

Aldiz, patroian zehazturiko baldintzetan, osagaiak HAULa ez beste konbinazioren batean ere agertzen ahal bada, HAUL hori anbiguoa dela esango dugu. Adibidez, baldin eta *bat* eta *egin* hitzak 12 ordenan agertzen badira, *egi*n osagaiak edozein

flexio hartzen duelarik, batzuetan *bat egin* aditz-lokuzioaren agerpenak izaten dira (26); beste askotan, ostera, bestelako konbinazio bati dagozkio (27).

(26) Eta, jendearen txaloek eta oihuek airean *bat egiten* zutela, turutots bat aditu zen (*Lur bat haratago*, Joan Mari Irigoien)

(27) Emilek buruz baiezko keinu *bat egin* zuen, serio-serio (*Manhattan Transfer*, J. Dos Passos / Lopez de Arana)

Ohar bedi HAUL bat ziurra izan daitekeela hurrenkera-eta flexio-baldintza jakin batzuetan, eta anbiguoa, aldiz, beste patroi batzuen arabera.

Bestalde, anbiguotasuna bi motatakoa izan daiteke. Batzuetan, hitzak patroian zehaztutako eran agertzen dira testuan, baina euren arteko erlazio sintaktikoa HAULaren osagaien artekoa ez bezalakoa da. Konparazio baterako, (27) eta (28) adibideetan *bat egin* eta *aldean beste* HAULen patroiak zehatz-mehatz betetzen diren arren, osagaiak sintagma bereizitan azaltzen dira (*baiezko keinu bat / egin zuen; ezkerreko aldean / beste horrenbeste*), eta ez dute, beraz, hitz anitzeko unitaterik osatzen.

(28) Ezkerreko *aldean beste* horrenbeste egin zion (*Fantasiazko ipuinak*, G. de Maupassant / Josu Zabaleta)

Beste batzuetan, (29) eta (30) adibideetan bezala, testuko hitzen arteko erlazio sintaktikoa eta HAULaren osagaien artekoa bat badatoz ere, agerpenak ez dagozkio HAUL interpretazioari.

(29) [...] eta sukaldeko burdina eskuan hartu, eta *kolpe batez* akabatu zuen katua (*Lur bat haratago*, Joan Mari Irigoien)

(30) Neskamearen atzetik zaunkaka joan zen etxeko *txakur txikia* (*101 gau*, "Txiliku")

Kasu hauetan, hitzok euren zentzu literalean erabiliak izan dira, inolako idiomatikotasun aztarnarik gabe, eta beraz ez dute unitate lexikalik osatzen.

5.1.4. Adibideak

Adibide pare baten bidez saiatuko gara argitzen HAULak deskribatzeko adierazpideaz 5.1. atal honetan orain arte azaldutakoa. Horretarako, EDBL datu-basetik *a bildua* eta *lan egin* sarreren informazioa esportatu dugu modu irakurgarrian 2. eta 3. irudietan hurrenez hurren.

```
"a bildua"   0    IZE ARR
    a        3    IZE ARR
    bildu    0    ADI SIN    MORF
    12    ZIUR
        a         [-]
        bildua    {/ADM=PART/}
```

2. irudia: a *bildu*a HAULaren gauzatze-eskema.

a bildua sarreraren deskripzioan (2. irudia) hauxe irakur dezakegu:

- Lehenengo lerroan, HAULaren lema kanonikoa (`"a bildua"`) eta homografo-identifikatzaileaz gain (`0`)[5], kategoriaz izen arrunta dela adierazten da (`IZE ARR`).
- Hurrengo bi lerroetan, HAULaren osagaiak dagozkien sarrera bakunekin lotzen dira: lehenengo osagaia *a* izen arruntarekin (`a 3 IZE ARR`) eta bigarrena *bildu* aditzarekin (`bildu 0 ADI SIN`). Halaber, HAULak bere informazio morfologikoa *bildu* osagaitik hartzen duela ere zehazten da (`MORF`).
- Ondoren, gauzatze-eskema deskribatzen da. Hasteko, hurrenkera eta anbiguotasuna adierazten dira. Kasu honetan, HAULaren osagaiak ordenaturik eta jarraian (elementu tartekaturik gabe) azal daitezke (`12`), eta hurrenkera horretan HAULa ez da anbiguoa (`ZIUR`).
- Hurrengo lerroetan, osagai bakoitzaren flexio-murriztapenak azaltzen dira, alegia, a osagaiak ez du batere flexiorik hartzen (`[-]`) eta *bildua* osagaiak, aldiz, *bildu* aditzaren partizipioko aspektu-morfema har dezake bakarrik[6] (`{/ADM=PART/}`).

```
"lan egin"   0    ADI ADK
    lan      0    IZE ARR
    egin     1    ADI SIN    MORF
    1*2    ANBI
        lan       {(KAS=ABS ETA MUG=MG) EDO KAS=PAR}
        egin      [%]
    2*1    ANBI
        lan       {(KAS=ABS ETA MUG=MG) EDO KAS=PAR}
        egin      [%]
```

3. irudia: *lan egin* HAULaren gauzatze-eskemak.

Era berean, *lan egin* sarreraren kasuan (3. irudia):

- *egin* osagaiak ematen dio informazioa HAUL osoari (`MORF`).
- Bi gauzatze-eskema ditu: lehenengoa (`1*2`) hurrenkerari dagokio, alegia, osagaiak ordenaturik doaz, eta jarraian zein etenik ager daitezke (tartean zero hitz edo gehiago dutela). Ordena honetan, HAULa anbiguoa da (`ANBI`).

---

[5] Lema bera duten sarrera ezberdinak bereizteko erabiltzen den zenbakiari deritzo homografo-identifikatzailea (adib. *bero* izena eta *bero* izenondoa).

[6] Kasu honetan, lemak hartzen duen lehenengo morfema zehazten da bakarrik. Beraz, partizipioko aspektu-morfemaren atzean zer beste atzizki joan daitezkeen (*bildu*, *bildua*, *bilduarekin*...) aspektu-morfema horrexen jarraitze-klaseak zehazten du.

- Bigarren gauzatze-eskeman, osagaiek trukatua dute ordena eta jarraian zein etenik azal daitezke (`2*1`). Hurrenkera honetan ere, HAULa anbiguoa da (`ANBI`).
- Bi gauzatze-eskemetan, osagaien flexio-murriztapenak berdinak dira: *lan* osagaia absolutibo mugagabean zein partitiboan joan daiteke (`{(KAS=ABS ETA MUG=MG) EDO KAS=PAR}`), eta *egin* aditzak edozein flexio har dezake (`[%]`).

5.2. HABIL

HAULen tratamendua burutzeko, HABIL izeneko tresna inplementatu dugu. Tresna honek hitz anitzeko unitateok identifikatu eta dagozkien interpretazioak esleitzen dizkie, datu-base lexikalean eginiko deskripzioetan oinarriturik.

HABILen ezaugarri nagusiak honakoak dira:

- HAUL jarraituak zein etenak tratatzen ditu
- Osagaien ordena posible guztiak hartzen ditu kontuan
- Flexio-murriztapen guztiak betetzen direla egiaztatzen du
- HAULen interpretazio morfosintaktikoak sortzen ditu

Horretarako guztirako, analizatzaile morfosintaktikoak sorturiko hitz bakunen analisietatik abiatzen da (ikus 4. atala). Bi fasetan burutzen da prozesua. Lehenengo, motor bilatzaile batek (*searching engine*) HAULa identifikatzen du testuan, bere deskripzioa (barne-osaketa eta gauzatze-eskema) baliatuta. Ondoren prozesadore morfosintaktiko batek HAULari dagozkion interpretazioak sortzen ditu.

Interpretazioak eraikitzeko, batetik, HAULak berak EDBLn dituen lema kanonikoa eta kategoria/azpikategoria ezaugarriak eransten dizkio. Bestetik, —HAULak osagaietako baten informazio morfologikoa heredatu behar duenean— prozesadoreak patroi-parekatze (*pattern-matching*) teknikak aplikatzen ditu, osagaiaren analisietatik dagozkion ezaugarri morfologikoak erauzi, eta ezaugarriok HAULaren interpretazioan eransteko. Esate baterako, *lan egin* aditz-lokuzioak *egin* osagaitik hartzen du informazio morfosintaktikoa. Ostera, *batez ere* lokuzio fosilduak ez du informazio morfologiko gehigarririk behar, beste formarik onartzen ez duelako, eta aski du lema, kategoria eta azpikategoria informazioarekin.

Ondoren, lokuzioa ziurra bada, osagaiei dagozkien interpretazioak ezabatu eta HAUL osoari dagozkionak gehitzen dira. Aldiz, lokuzioa anbiguoa denean, hitz anitzekoaren interpretazioak *erantsi* egiten zaizkio osagaien analisiei.

4. irudian, *Erlijio alorrean Zaitegi-k berak egin zuen lan batez ere* esaldiaren analisia ikus dezakegu, testu-hitz bakoitzaren interpretazio guztiekin. Ikus dezakegunez, *egin* eta *lan* osagaien kasuan, *lan egin*-en interpretazioak gehitu egin zaizkio osagai gisa zituztenei, izan ere, *lan egin* lokuzioa, hurrenkera jakin horretan, anbiguo gisa dago deskribatua datu-basean (*Hainbat lan egin ditugu*). Aldiz, *batez ere* lokailuak analisi bakarra du, lokuzioari dagokiona, HAUL ziurra baita.

```
"<Erlijio>"<HAS MAI>"
        "erlijio" IZE ARR ZERO HAS MAI @KM
        "erlijio" IZE ARR ABS MG HAS MAI @OBJ @PRED @SUBJ
"<alorrean>"
        "alor" IZE ARR INE NUMS MUGM @ADLG
"<Zaitegi-k>"<HAS MAI>"
        "Zaitegi" IZE LIB ERG NUMS MUGM HAS MAI @SUBJ
"<berak>"
        "bera" ADJ ARR ERG MG AORG @SUBJ
        "bera" ADJ ARR ABS NUMP MUGM AORG @OBJ @PRED @SUBJ
        "bera" ADJ ARR ERG NUMS MUGM AORG @SUBJ
        "berak" DET ERKIND NMGP ABS NUMP MUGM @OBJ @PRED @SUBJ
        "bera" DET ERKIND NMGS ERG NUMS MUGM AORG @SUBJ
"<egin>"<1-2>"
        "egin" ADI SIN ADOIN NOTDEK @-JADNAG
        "egin" ADI SIN PART BURU NOTDEK @-JADNAG
        "egin" IZE ARR ABS MG @OBJ @PRED @SUBJ
        "lan egin" ADI ADK ADOIN NOTDEK @-JADNAG
        "lan egin" ADI ADK PART BURU NOTDEK @-JADNAG
        [...]
"<zuen>"
        "*edun" ADL B1 NOR NORK NR HURA NK HARK @+JADLAG
        "ukan" ADT PNT B1 NOR NORK NR HURA NK HARK @+JADNAG
        "zuek" IOR PERARR ZUEK GEN NUMP MUGM ZERO @<IZLG @IZLG>
        [...]
"<lan>"<2-2>"
        "landu" ADI SIN ADOIN NOTDEK @-JADNAG
        "lan" IZE ARR ABS MG @OBJ @PRED @SUBJ
        "lan egin" ADI ADK ADOIN NOTDEK @-JADNAG
        "lan egin" ADI ADK PART BURU NOTDEK @-JADNAG
        [...]
"<batez_ere>"
        "batez ere" LOT LOK EMEN @LOK
"<$.>"<PUNT PUNT>"
        PUNT PUNT
```

4. irudia: *Erlijio alorrean Zaitegi-k berak egin zuen lan batez ere* esaldiaren analisia.

5.3. Desanbiguaziorako murriztapen-gramatika

Arestian esan dugun moduan, HAULa anbiguoa denean, HABIL tresnak hitz anitzeko unitateari dagozkion interpretazioak eransten ditu, osagaiei dagozkienak ezabatu gabe. Une horretan, HAUL-hautagaiak besterik ez dira. Bada, hautagaiok zein testuingurutan diren egiaz hitz anitzeko unitate eta zeintzuetan ez diren ebatziko duen gramatika bat garatzen ari gara hurrengo urrats honetan.

Gramatika egiteko, *Murriztapen Gramatika*[7] (MG) formalismoa (Karlsson *et al.*, 1995; Voutilainen eta Tapanainen, 1993; Tapanainen, 1996) baliatu dugu. MG egoera finituko prozesadore sendoa da, eta bere egitekorik behinena desanbiguazioa du, hau da, testu-hitzen irakurketa desegokiak baztertzen joatea harik eta interpretazio zuzen bakarrarekin (edo ahal den gutxienekin) gelditu arte. Horretarako, erregelak idazten dira, testuinguruko informazioa erabiliz. 1990ean lehenengoz merkaturatu zenetik, hainbat hizkuntza motatarako analizatzaile morfologikoak zein sintaktikoak garatu dira MG erabiliz, emaitza zinez onekin.

Gramatikaren ondorengo erregelak, adibidez, HAUL gisa markatuko ditu *hitz egin* erako aditzak *egin zuen hitz* edo *egin zezakeen hitz* bezalako segidetan.

```
ADD (%HAUL_BAI) TARGET HITZ-EGIN IF
(0 EGIN)
(1 EDUN/EZAN) (NOT 1 NOR-HAIEK)
(2 HITZ AND HITZ-EGIN)
(NOT 3 IZENONDO OR POSDET) ;
```

Erregela honek `%HAUL_BAI` etiketa ezarriko dio (`ADD`) *hitz egin* lokuzioaren (`TARGET HITZ-EGIN`) *egin* osagaiari (`0 egin`), baldin eta:

- hurrengo hitza eskuinera **edun* edo **ezan* aditzaren formaren bat bada (`1 EDUN/EZAN`), baina nor eremua haiek balioa ez badu (`NOT 1 NOR-HAIEK`)
- 2. hitza lokuzioaren *hitz* osagaiari badagokio (`2 HITZ AND HITZ-EGIN`) eta
- 3. hitza ez bada izenondoa edo izen osteko determinatzailea (`NOT 3 IZENONDO OR POSDET`)

Ikerlanaren une honetan, *XX. Mendeko Euskararen Corpus Estatistiko*ko 20 HAUL anbiguorik maizkoenak desanbiguatzeko gramatika garatu dugu. Horretarako, 110 erregela sortu ditugu guztira. Erregela multzo bera, jokaera berdina duten aditz-lokuzio guztiei aplikatzen zaie. Adibidez, *hitz egin* aditz-lokuzioarentzat garaturiko erregelak *lan egin* eta *ihes egin* aditzei ere aplikatzen zaie[8], eta aurrerago, mota bereko aditz guztiei aplikatuko zaie.

---

[7] Ingelesez *Constraint Grammar*

[8] 20 HAUL anbiguorik maizkoenen artean daude *la*n *egi*n eta *ihe*s *egi*n ere

5.4. Ondorioak

Hizkuntzan oro har, eta hizkuntzaren prozesamenduan partikularki, hitz anitzeko unitateek duten berealdiko garrantziaz ohartarazi nahi izan dugu artikulu honetan. Unitate fraseologikoen mota ezberdinei gainbegiratua egin diegu, eta gure lehentasuna lokuzioak tratatzea dela adierazi dugu. Horien prozesaketa automatikoa egiteko dauden hurbilpenak labur-labur aztertu ondoren, geure eredua azaldu dugu: HAULen deskripziorako adierazpide bat, horiek detektatzeko eta analizatzeko HABIL tresna eta HAUL anbiguoen desanbiguaziorako gramatika.

Bide luzea dago oraindik zeregin azkengabe honetan. Deskribaturiko unitate konplexuen kopurua txikia da beste hizkuntza batzuetakoen aldean. Desanbiguaziorako gramatika ere oso mugatua da oraindik, eta gainera, ebaluatzeko dago.

Beraz, eman beharreko hurrengo pausoak hauexek lirateke. Oso epe laburrean, egindako gramatika ebaluatuko dugu. Hurrenik, EDBLn deskribaturik dauden HAUL anbiguoak sailkatuko ditugu, jokaera berdina dutenak multzokatuz. Horietako batzuei —20 HAUL maizkoenen multzokoak izanik— eginda dauden desanbiguazio-erregelak aplikatu ahalko zaizkie; gainontzeko multzoetarako, ostera, erregela berriak sortu beharko dira.

Aurrerago, datu-basea aberasten eta desanbiguazio-gramatika eguneratzen jarraitu behar dugu etengabe.

Erreferentziak

Aduriz I. EUSMG: Morfologiatik sintaxira Murriztapen Gramatika erabiliz. Euskararen desanbiguazio morfologikoaren tratamendua eta azterketa sintaktikoaren lehen urratsak. Doktoretza-tesia, Filologia eta Historia-Geografia Fakultatea. UPV-EHU, Gasteiz, 2000.

—; Agirre E.; Aldezabal I.; Alegria I.; Ansa O.; Arregi X.; Arriola J.M.; Artola X.; Díaz de Ilarraza A.; Ezeiza N.; Gojenola K.; Maritxalar M.; Oronoz M.; Sarasola K., Soroa A. eta Urizar R. A framework for the automatic processing of Basque. *Proceedings of Workshop on Lexical Resources for Minority Languages*, Granada, Spain, 1998.

—; —; —;—; Arregi X.; Arriola J.M.; Artola X.; Gojenola K.; Maritxalar M.; Sarasola K. eta Urkia M. A word-grammar based morphological analyzer for agglutinative languages. *COLING, 18th International Conference on Computational Linguistics*, 1–7, Universität des Saarlandes, Saarbrücken, Germany, 2000. Morgan Kaufmann.

—; Aranzabe M.J.; Arriola J.M.; Díaz de Ilarraza A.; Gojenola K.; Oronoz M., eta Uria L. A cascaded syntactic analyser for Basque. In Gelbukh A., editor, *Computational Linguistics and Intelligent Text Processing: 5th International Conference CICLing2004*, 2945 lib. of *Lecture Notes in Computer Science*, 124–134. Springer-Verlag GmbH, February 15-21 2004. ISBN 3-540-21006-7.

Alegria I. *Euskal morfologiaren tratamendu automatikorako tresnak*. Doktoretza-tesia, Informatika Fakultatea. UPV-EHU, uztaila 1995. Kepa Sarasola eta Xabier Artola, UPV-EHUko irakasleen zuzendaritzapean eginiko tesia.

Arriola J.M.; Esparza I.; Ezeiza N.; Gojenola K. eta Sologaistoa A. *Analizatzaile morfosintaktikoa*. Barne-txostena, EHU/UPV, 2005.

Beesley K.R. eta Karttunen L. *Finite State Morphology*. CSLI Publications, 2003.

Breidt E.; Segond F. eta Valetto G. Formal description for multi-word lexemes with the finite-state formalism IDAREX. *Proceedings of COLING-96, Copenhagen,* 1036–1040, 1996.

Copestake A.; Lambean F.; Villavicencio A.; Bond F.; Baldwin T.; Sag I. eta Flickinger D. Multiword expressions: linguistic precision and reusability. *Proceedings of the Third International Conference on Language Resources and Evaluation. LREC*, 1941–1947, Las Palmas, 2002.

Corpas G. *Manual de Fraseología Española*. Editorial Gredos, Madrid, 1996.

Courtois B. eta Silberztein M. *Les dictionnaires électroniques du français*, 87 lib. Larousse, Langue française, 1990.

Euskaltzaindia. *Euskera-45*, Hiztegi Batua. Euskaltzaindia, Bilbo, 2002.

Ezeiza N. Corpusak ustiatzeko tresna linguistikoak. *Euskararen etiketatzaile sintaktiko sendo eta malgua*. Doktoretza-tesia, EHU/UPV, Donostia, 2002.

Gojenola K. *Euskararen sintaxi Konputazionalerantz. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan*. Doktoretza-tesia, Informatika Fakultatea. Euskal Herriko Unibertsitatea, Donostia, 2000.

Jackendoff R. *The architecture of the language faculty*. Cambridge, MA: MIT Press, 1997.

Jacquemin C. *Spotting and discovering terms through Natural Language Processing*. MIT Press, 2001.

Karlsson F.; Voutilainen A.; Heikkila J. eta Anttila A. *Constraint Grammar: Language-independent System for Parsing Unrestricted Text*. Prentice-Hall, Berlin, 1995.

Karttunen L. *Finite-state lexicon compiler*. Barne-txostena, Xerox PARC, 1993.

—; Kaplan R.M. eta Zaenen A. Two-level morphology with composition. *Proceedings of COLING-92, Nantes,* 141–148, 1992.

Koike K. *Colocaciones léxicas en el español actual: estudio formal y léxico-semántico*. Universidad de Alcalá / Takushoku University, 2001.

Oflazer K. Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. *Computational Linguistics*, 22(1):73–89, 1996. ISSN 0891-2017.

Sag I.A.; Baldwin T.; Copestake F.B.A. eta Flickinger D. Multiword expressions: a pain in the neck for NLP. *CICLing-2002*, *Mexico*, 1–15, 2002.

Savary A. Computational inflection of multi-word units. a contrastive study of lexical approaches. *LiLT*, 1 Issue 2:1–53, 2008.

Shieber S.M. *An Introduction to Unification-Based Approaches to Grammar*. Number 4. CSLI Lecture notes, Stanford, 1986.

Silberztein M. *Dictionnaires électroniques et analyse automatique de textes: Le système INTEX*. Masson, Paris, 1993.

Tapanainen P. *The Constraint Grammar parser CG-2*. Publications of the University of Helsinki, 27, Helsinki, 1996.

Urkia M. *Euskal morfologiaren tratamendu informatikorantz*. Doktoretza-tesia, Filologia eta Historia-Geografia Fakultatea. UPV-EHU, uztaila 1997.

Villavicencio A.; Copestake A. eta and Fabre Lambeau B.W. Lexical encoding on mwes. *Second ACL workshop on Multiword Expressions: Integration Processing*, 80–87, 2004.

Voutilainen A. eta Tapanainen P. Ambiguity resolution in a reductionistic parser. *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, 394–403, Morristown, NJ, USA, 1993. Association for Computational Linguistics. ISBN 90-5434-014-2.

Zabala I. *Predikazioaren teoriak gramatika sortzailean (euskararen kasua)*. Doktoretza-tesia, EHU/UPV, Gasteiz, 1993.

Zabala I. *Las fronteras de la composición en lenguas románicas y en vasco*. Los predicados complejos en vasco, 445–534. Universidad de Deusto, 2004.