EuskalWordNet: euskararako ezagutza-base lexiko-semantikoa

Eneko Agirre, Izaskun Aldezabal, Jone Etxeberria, Eli Izagirre, Karmele Mendizabal, Eli Pociello and Mikel Quintian*

IXA taldea
Euskal Herriko Unibertsitatea
e.agirre@ehu.es; izaskun.aldezabal@ehu.es; elisabete@si.ehu.es

Abstract

Natural Language Processing techniques need to develop lexical-semantic knowledge bases (LSKB) in order to perform semantic interpretation. The IXA group decided to develop a Basque LSKB called EuskalWordNet for this reason. EuskalWordNet is based on WordNet and its multilingual counterpart EuroWordNet. This paper reviews the theoretical and practical aspects of the EuskalWordNet LSKB, as well as the steps followed in its construction.

Laburpena

Lengoaia Naturalaren Prozesamenduan semantika jorratu ahal izateko ezinbestekoak dira ezagutza-base lexiko-semantikoak (EBLSak) garatzea. Arrazoi horregatik, badira urte batzuk IXA taldea euskararako EBLS bat garatzen hasi zela, EuskalWordNet deiturikoa. EuskalWordNet WordNet-en eta honen ildotik garatutako EuroWordNet-en oinarritua dago. Artikulu honetan, IXA taldeak EBLS bat egiteko jarraitutako pausoak azaltzeaz gain, EuskalWordNet EBLSaren alderdi teoriko eta praktikoak deskribatzen dira.

Keywords: natural language processing, linguistic resources, lexical-semantic knowledge base, computational semantics.

Hitz gakoak: lengoaia naturalaren prozesamendua, hizkuntza baliabideak, ezagutza-base lexiko-semantikoak, semantika konputazionala.

1. Sarrera

Euskal Herriko Unibertsitateko Informatika Fakultateko IXA taldeak hamar urte baino gehiago daramatza Lengoaia Naturalaren Prozesamenduan (LNP) lanean. Arlo zabal horren barruan, euskararen gaineko ikerketa aplikatua da gure xede nagusia, eta helburu horrekin orain arte, batez ere, morfologia (Aduriz et al., 1994, besteak beste) eta sintaxia (Aduriz et al., 1998, besteren artean) landu ditugu. Arlo hauetan lan handia egiteke dagoen arren, hurrengo aurrerapauso garrantzitsua semantika jorratzea da.

Semantika beharrezkoa da hainbat ataza konputazionaletan aurrera egin ahal izateko (egitura sintaktikoen desanbiguazioan, hitzen adieren desanbiguazioan, anaforaren ebazpenean, itzulpen automatikoan…). Arrazoi horregatik, IXA taldean dagoeneko hasiak gara ezagutza lexiko-semantikoaren ikasketan murgiltzen, esate baterako euskarako aditzen alternantzien eta klase semantikoen azterketan (Aldezabal, 2004), aditzen adieraren desanbiguazioan (Martínez, 2005), eta abar.

Lan hauei guztiei etekin handiagoa aterako litzaieke formalismo beraren arabera antolatutako ezagutza-base batean egongo balira. Azpimarratzekoa da, bestalde, betebehar hauek ez direla IXA taldearenak bakarrik, semantika konputazionala edo hizkuntzaren ulermena burutu nahi duen edozein talderenak baizik, egun semantika konputazionalaren arloan zein hizkuntzaren inguruan sortzen ari den industrian, ezinbestekoak baitira baliabide lexikalak; besteak beste, hitzen esanahia emango duten baliabideak.

Hori dela eta, IXA taldean informazio lexiko-semantikoa jasotzen duen ezagutza-base lexiko-semantiko (aurrerantzean EBLS) bat garatzen ari da: EuskalWordNet.

2. EBSLak

Lengoaia naturalen prozesamendu sintaktiko eta semantikoa egin ahal izateko, lexikoiak hitz-zerrenda izatetik EBLS izatera pasatu dira, hitzei eta adierei buruzko informazioa dutenak. EBLS batean, hizkuntza ulertu ahal izateko, ordenagailuan hitzei buruz jakin beharreko guztia egon beharko litzateke.

EBLSak baliabide lexikal egituratuak ditugu, hitzei eta adierei buruzko informazioa dutenak, alegia. Esaterako, EBLS asko hierarkikoki antolatuta daude (baita EuskalWordNet bera ere).

Historikoki, baliabide lexikalak eskuz egiten ziren; baina, informazio-kopuru itzela landu behar zela-eta ahalegin handia eskatzen zutela kontuan izanik, laguntza automatiko eta erdiautomatikoen bidea jorratzeari ekin zaio azken hamarkadan.

(*) Autoreak alfabetikoki izendatuak daude.

2.1. EBLSa definitzeko zailtasunak

EBLS bat lantzeko orduan zenbait zailtasun topa daitezke. Batetik, EBLSak egiteko eredu edo formalismoen aniztasuna dago. Ondorioz, hizkuntzalaritza teorikoan eredu ugari proposatu izan dira, (Dowty, 1979; Jackendoff, 1990; Talmy, 1985, besteak beste) baina beraien artean ez dago batasunik, eta batzuetan gainera, bata bestearekin kontraesanean daude.

Hizkuntzalaritza konputazionalean ere proposamenak ugariak dira (Bresnan eta Kaplan, 1982; Fillmore eta Baker, 2001; Miller, 1985; Kipper et al., 2000, eta abar), baina askotan fenomeno linguistiko zehatz bati mugatutako EBLSak dira.

Bestetik, definitzen zailak diren fenomeno linguistikoak zehaztu behar dira. Esaterako, ale lexikalak definitzen dituen EBLS bat sortzerakoan, ale lexikalak semantikaren ikuspegitik ere definitzen ari garela pentsatu behar da. Hala ere, ez dago batere argi ale lexikal batek izan behar dituen ezaugarri semantikoak zeintzuk diren. Izan ere, egun oraindik iritzi ezberdinak daude ale lexikalen izaera semantikoa definitzerakoan: ale lexikalak berezko semantika du ala testuinguru sintaktikoaren eraginaren ondorioz jaso du semantika hori? Eta hori horrela izanda, zer ezaugarri dira ale lexikalean berezkoak eta zeintzuk dira testuinguru sintaktikoaren eraginaren ondorioz sortutakoak?

Horrela bada, EBLSak ale lexikalen izaera semantikoa definitzerakoan zenbait ikuspegi izan ditzake: semantiko hutsa, sintaktikoa edo sintaktiko-semantikoa. Hortaz, EBLSaren ikuspegiaren arabera sarrera lexikala ezaugarri desberdinekin zehaztua etorriko da.

2.1.1. Euskararako EBLSaren eredu aukeraketa

Gorago ikusi dugun bezala (2.1 atalean), EBLSen eraikuntzarako ez dago eredu bat bakarrik, ez hizkuntzalaritza teorikoan ezta konputazionalean ere. Proposamen ugari daude, eta hizkuntzalaritza konputazionalaren kasuan, proposamen hauek arloetan zehar sakabanatuak daude. Euskararako EBLS bat egiten hasi baino lehen, zenbait eredu edo formalismo aztertu dira. LNPren arloan jorratuak izan direnak interesatu zaizkigu bereziki –FrameNet (Fillmore eta Baker, 2001), WordNet (Miller, 1985; Fellbaum, 1998), EuroWordNet (Vossen, 1998), Volem (Fernández et al., 2002)–, baina askotan hauek lan teorikoetan oinarrituak daudenez, garrantzitsua iruditu zaigu lan teoriko hauen ezagutza ere izatea: Jackendoff (1990), Levin (1993), Pustejovsky (1995) eta abar.

Azkenik, euskarako EBLSaren diseinua irizpide batzuetara mugatu dugu, hau da, euskarako ezagutza-basea egiteko aukeratzen genuen EBLS ereduak ondorengo baldintzak betetzea nahi genuen:

- Hizkuntza bere osotasunean adierazten duen EBLSa izatea, ale lexikal bakoitzari dagokion adiera, klase semantikoa eta informazio sintaktiko-semantikoa zehaztuta dituen EBLSa.
- Ahal dela, teoria edo ikerlan bakar bati lotua ez dagoen EBLS eredua izatea, hau da, beste eredu edo formalismo batzuetatik edan dezakeen EBLSa izatea.
- Konputazionalki inplementa daitekeen EBLSa izatea, hots, LNPn erabilgarria dena.
- Aukeratutako eredu horretatik gertu beste lan konputazionalak egotea, gure EBLSa horien informazioarekin ere aberastu ahal izateko.

2.2. Aukeratutako EBLS eredua

Aurreko atalean aipatutako azterketa horren ondorio gisa, eta aipatu ditugun irizpideen arabera, IXA taldearen beharretara gehiago egokitzen den EBLS formalismoa WordNet eta honen ildotik abiatuta garatu den EuroWordNet direla esango dugu[1].

WordNet (Miller, 1985; Fellbaum, 1998) teoria psikolinguistikoetan oinarritua dagoen ingeleseko EBLSa da. Princeton-eko Unibertsitatean garatzen ari da –Cognitive Science Laboratory delakoan– George A. Miller-en ardurapean.

Ingeleseko izen, aditz, adjektibo eta adberbioak *synonym set* edo *synset*-etan (sinonimo multzotan) antolatuak daude, hauetako bakoitza kontzeptu lexikal bati dagokiolarik. Esaterako, ingeleseko *T-shirt* izenak WordNet-en hurrengo *synset*-a (adiera[2]) du, euskaraz ‘elastiko’ izendatzen duguna (ikus 1. irudia):

T-shirt, jersey *-- (a close-fitting pullover shirt)*

1. Irudia: *T-shirt* izenaren *synset*-a WordNet-en

Synset-a bi ale lexikalez osatua dago (*T-shirt*, eta *jersey*), hots, *T-shirt* eta *jersey* izenak *synset* horretan sinonimoak dira.

Sinonimiaz gain WordNet-eko *synset*-en artean, erlazio lexikal anitz daude. Garrantzitsuena hiperonimia-hiponimia erlazioa dela esan daiteke, hau baita *synset*-ak hierarkiatan eta multzo semantiko nagusietan multzokatuko dituena. Honi buruz sakonago arituko gara EuskalWordNet azaltzerakoan (ikus 3. atala).

WordNet edozeinek eskura dezake Internet bidez[3], eta gaur egun oso erabilia da LNP inguruko ikerkuntzan[4].

[1] Argibide gehiago Pociello (2004) lanean.
[2] Aurrerantzean *synset* terminoa erabiliko dugu.
[3] www.cogsci.princeton.edu/cgi-bin/webwn

EuroWordNet proiektua (Vossen, 1998) 1996an hasi zen proiektu europarra da. Ezagutza-base eleanitza da, Europako zortzi hizkuntzatara zabaltzen dena (ingelesa, daniera, italiera, gaztelania, alemana, frantsesa, txekiera eta estoniera). EuroWordNet-ek Princeton-eko WordNet-aren eredua jarraitzen du, hots, Princeton-en ingeleserako egindako WordNet-aren *synset*, harreman semantiko eta hierarkian oinarritu dira beraien WordNet-a sortzeko. Hala ere, WordNet-en egitura, harreman eta *synset*-etan oinarritu arren, WordNet-ek ez zituen ezaugarri batzuk EuroWordNet-en gaineratu dira: oinarrizko kontzeptuak, domeinu- eta goi-ontologiak, besteak beste[5].

Nahiz eta EuroWordNet-en hizkuntza bakoitzak WordNet "independente" bat izan, EuroWordNet-en helburua WordNet desberdin hauek guztiak ezagutza-base eleanitz bakarrean elkartzea da. Horretarako, hizkuntza guztien WordNet guztiak elkargune bat dute, *Inter-Lingual-Index*-a (hemendik aurrera ILI) deritzana, aldi berean, Princeton-eko WordNet 1.5 bertsioari lotua dagoena. ILI honen bitartez, hizkuntza guztietako WordNet-ak lotuak daude, eta ingeleseko *synset*-a EuroWordNet-a osatzen duten hizkuntza guztietan ikusgarri egongo da. Beste hitz batzuetan esanda, *synset* bera ingelesez, danieraz, italieraz, gaztelaniaz, alemanez, frantsesez, txekieraz eta estonieraz agertzen da. 1. irudiko WordNet-eko *synset* bera, 2. irudian dugu EuroWordNet-eko interfazean ingelesez eta gaztelaniaz (glosa ingelesez bakarrik dator). Hizkuntza hauetako guztietako ordainak, 02874798n ILIaren bidez lotzen dira.

jersey_1 T-shirt_1 a close-fitting pullover shirt
02874798n
camiseta_1 niqui_1

2. Irudia: *T-shirt* izenaren *synset*-a ingelesez eta gaztelaniaz EuroWordNet-en.

2.2.1. Aukeratutako EBLS eredutik nola abiatu

Behin euskarako EBLS egiteko oinarrituko garen eredua erabakita, eta ikusita EBLS hori ingeleserako sortu dela (aztertutako EBLS gehienak bezala), beste erabaki berri baten aurrean gaude: euskaraz dauden corpus eta hiztegietatik abiatuta euskarako EBLSa sortu, ala euskararako EBLSa egitea, erdararako egin diren EBLSez baliatuta.

Lehenengo aukeran, sortu beharreko adierak eta hierarkiak WordNet-eko hierarkiekiko independente izango lirateke. Baina, hurbilpen horrek lan lexikografiko handia eskatuko luke, eta, horrez gain, hizkuntzen arteko adieren loturak adierazteko bideak sortu beharko lirateke. Bigarren aukeran, WordNet abiapuntu gisa hartuz gero, nahiz eta guk ez kontrolatu adieren sorkuntza eta antolamendu hierarkikoa, dohainik dugu ingelesezko kontzeptuekiko lotura, eta hizkuntzen arteko adieren loturak egiteko bidea ere ematen zaigu (ILIaren bidez).

Bi hurbilpenen alde onak eta txarrak aztertu ondoren, euskararako Wordnet-a egiteko EuroWordNet abiapuntutzat hartzea erabaki dugu –EuskalWordNet (Agirre et al., 2002)–, hau da, EuroWordNet-eko ingelesezko kontzeptuei euskarazkoak lotuz. Horregatik, hurrengo ataletan EuskalWordNet-en ezaugarriak azaltzean, aldi berean, EuroWordNet-en ezaugarriak deskribatzen arituko gara.

3. EuskalWordNet

3.1. EBLSaren antolaketa

Aipatu dugun bezala, EuskalWordNet WordNet eta EuroWordNet-en ildotik garatzen ari garen EBLSa da. Hala, EuskalWordNet *synset*-en (sinonimo multzoen) arabera antolatua dagoen hierarkia kontzeptuala da. 1. eta 2. irudiak gogora ekarriz, 3. irudian euskarako *elastiko* izenaren *synset*-a aurkezten dugu. Kasu honetan, ingeleseko glosarekin batera, euskarazkoa ere landuta dago.

jersey_1 T-shirt_1 a close-fitting pullover shirt
02874798n camiseta_1 niqui_1
elastiko_1 kamiseta_1 niki_1 puntuzko jantzia, leporik ez duena

3. Irudia: *elastiko* izenaren *synset*-a ingelesez, gaztelaniaz eta euskaraz EuskalWordNet-en.

3. irudiko *synset*-ean beste hizkuntzetako ordainak ikusgarri ditugu; denek kontzeptu bera adierazten dute –'puntuzko jantzia, leporik ez duena'– eta horregatik denek ILI zenbaki bera daramate: 02874798n.

EuskalWordNet-eko erlazio semantiko garrantzitsuenetako bat sinonimia dela nabarmena da; ezagutza-basearen oinarria ale lexikalaren adieran baitago, eta adiera hori ale lexikal batek baino gehiago duenean, ale lexikalak multzokatu egiten direlako.

Hala ere, EuskalWordNet ez da *synset* zerrenda hutsa; *synset*-ak erlazio semantikoen bidez antolatuak daude. Esan dugun bezala, sinonimia da erlazio semantiko garrantzitsuenetakoa, baina honekin batera, beste hainbat erlazio daude, hala nola, hiperonimia-hiponimia erlazioa.

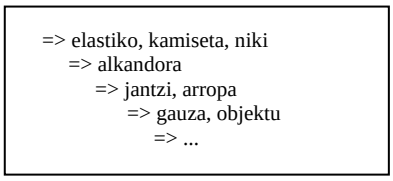
Hiperonimia-hiponimia erlazioak *synset* orokorrenak *synset* zehatzagoekin lotzen ditu[6]. 4. eta 5.

[4] Artikuluen zerrenda bat ikus daiteke WordNet-eko amaraun-orrian.

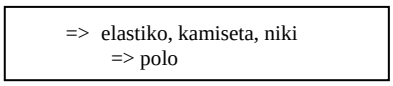
[5] EuskalWordNet EuroWordNet-en oinarrituta garatzen ari denez, 3. atalean EuskalWordNet azaltzerakoan aipatuko ditugu ezaugarri hauek guztiak (ikus 3.5.2 atala).

[6] Ingelesez *IS-A relation* bezala ere ezagutzen da, hots, *x is a kind of y*.

irudietan, 3.aren hiperonimoak eta hiponimoak ikus ditzakegu hurrenez hurren[7]:



4. Irudia: *elastiko* izenaren hiperonimoak EuskalWordNet-en.

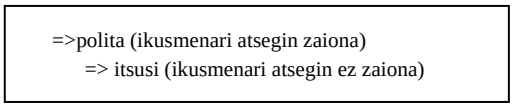


5. Irudia: *elastiko* izenaren hiponimoa EuskalWordNet-en.

4. irudian *elastiko* izenaren hiperonimoak ditugu, eta hauek *synset* horren ezaugarriak definitzen dituzte hierarkikoki. Esaterako, *elastikoa alkandora* mota bat bezala adierazten da; *alkandora jantzi* mota bat bezala eta aldi berean, *jantzia objektu* bat bezala. Ondorioz, *elastiko* izena, adiera honekin, *alkandora*, *jantzi*, eta *objektu* mota bat izango da.

Hiponimoak hiperonimoen zehaztapenak dira. Hortaz, 5. irudian *elastiko* izenaren adiera horretan, zehaztapen gisa elastiko motak agertzen dira (*polo*).

Horrela bada, EuskalWordNet ontologia edo hierarkia bat da, eta hiperonimia-hiponimia harreman semantikoarekin hierarkian gora eta behera egiteko aukera dugu. Ontologia hau kategoriaka banatua dago, eta kategoria bakoitzak bere hierarkia du; hau da, kategoria bakoitzaren hierarkia antolatzen da erlazioa semantiko nagusi baten arabera. Izen eta aditzen kasuan erlazio semantiko nagusia hiperonimia-hiponimia da[8]. Adjektibo eta adberbioek, berriz, sinonimia-antonimia dute ardatz gisa beraien antolakuntzan. 6. irudian, *polita* adberbioaren antonimoa ikus dezakegu (*itsusi*):



6. Irudia: *polita* adjektiboaren antonimoa EuskalWordNet-en.

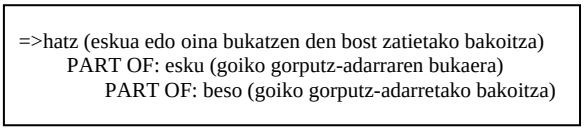
EuskalWordNet-eko sailkapena, beraz, *synset*-etan eta beraiek harremanetan jartzen dituzten erlazio semantikoetan datza. Erlazio semantiko hauen bidez, *synset*-ak multzokatzen dira, edo beste era batera esanda, klase semantikoak osatzen dira. *Synset* orokorragoren azpian (adabegi horren azpian) bere zehaztapenak multzokatzen dira. Esaterako, elastiko mota desberdinak *synset* baten azpian jasota daude (02874798n *synset*-aren azpian, alegia), hortaz, elastiko motak jasotzen dituen klase semantikoa *synset* horren bitartez adieraz daiteke.

3.2. Bestelako erlazio semantiko batzuk

Sinonimia eta hiperonimia-hiponimia erlazio semantikoetaz gain, EuskalWordNet-ek beste asko landu ditu. Hemen batzuen aipamen laburra egingo dugu[9].

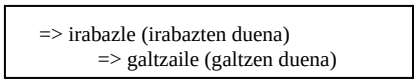
Izenak lotuak egon daitezke ***antonimia*** eta ***part-whole relations*** deituriko erlazio semantikoekin, besteak beste.

Part-whole relations deiturikoak: Zatia eta osotasuna harremanetan jartzen dituen erlazioak dira. Batetik, meronimia dago, *x is part of y* definizioari jarraitzen diona. Adibidez, *hatza eskuaren* zati bat da, eta *eskua*, aldi berean, *besoarena*:



7. Irudia: meronimia EuskalWordNet-en.

Antonimia: Izen batzuek antonimoak dituzte eta erlazio semantiko honek lotzen ditu:



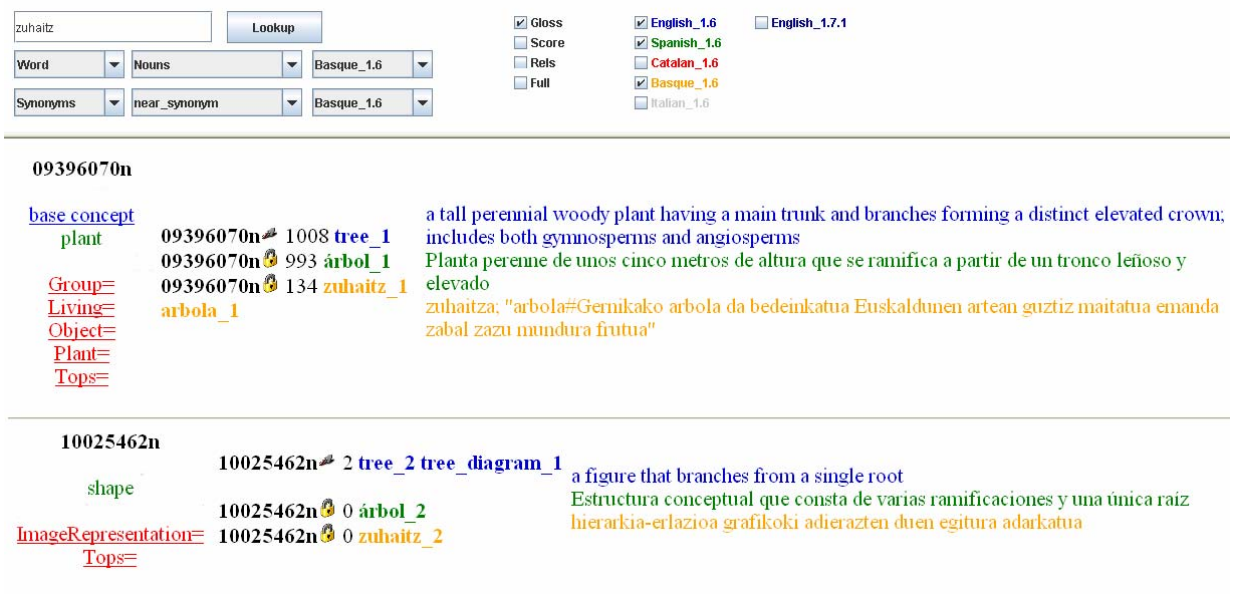
8. Irudia: antominia EuskalWordNet-en.

Aditzen hierarkian erlazio semantiko nabarmenetako bat ***entailment*** deritzona da: *verb1 logically entails verb2*. Esaterako, norbait zurrunka egiten ari bada, nahitaez, lo dago.

[7] Leku arazoak direla-eta, adibideak ezin izan ditugu osoak aurkeztu. Hiperonimo/Hiponimo hierarkia osoak ikusteko jo http://ixa2.si.ehu.es/mcr/wei.html web gunera.

[8] Aditzen kasuan, hiperonimia-troponimia erlazioz hitz egiten da. Argibide gehiago Fellbaum-en lanean (1998).

[9] Argibide gehiago (Fellbaum, 1998) eta (Miller, 1985) lanetan.



9. Irudia: EuskalWordNet-en interfazea.

Adjektibo eta adberbioen kasuan, erlazio semantiko gutxiago daude. Adjektibo batzuk (adibidez *eder*) adiera berdineko izenekin (*edertasun*) lotu egiten dira.

Esan bezala, erlazio semantiko batzuk baino ez ditugu aipatu. WordNet-en gehiago daude eta hauen kopurua handituz doa.

3.3. Metodologia

Ezaugarri orokorrak aipatu ondoren, EuskalWordNet-en garapenaren metodologia azalduko dugu. Metodologian hiru atal nagusi bereiz daitezke:

- Lehenengo urratsak, oinarrizko EuskalWordNet eraikitzea izan zuen xede. Euskarazko ordainak ingelesezko oinarrizko kontzeptuei (*Base Concepts* delakoei[10]) eskuz lotzea.
- Bigarren urratsean, ingelesezko *synset*-en euskal ordainak hiztegi elebidunak baliatuz (euskara-ingelesa[11]) automatikoki sortu ziren.
- Hirugarren urratsa, automatikoki sortu diren euskarazko *synset* horien kontzeptuz kontzeptuko eskuzko orrazketan datza.

Bestalde, garrantzitsua da azpimarratzea, EuskalWordNet-en garapenean oinarri eta abiapuntu gisa, ingelesezko WordNet 1.6 hartu dugula; beraz, gerta liteke euskarazko hainbat kontzeptu (*sagardotegi*, *ertzaina*...) lekurik ez izatea ingelesezko kontzeptuen artean. Horrelako "euskal kontzeptuei" dagokien *synset*-a sortu behar zaie eta hierarkian txertatu adabegi egokiaren azpian. Hau eskuzko orrazketa amaitzean, egiten hasi beharreko dagoen ataza da.

[10] Ikus 3.5.2 atala.
[11] Morris, 1998.

3.4. Egoera eta erabilera

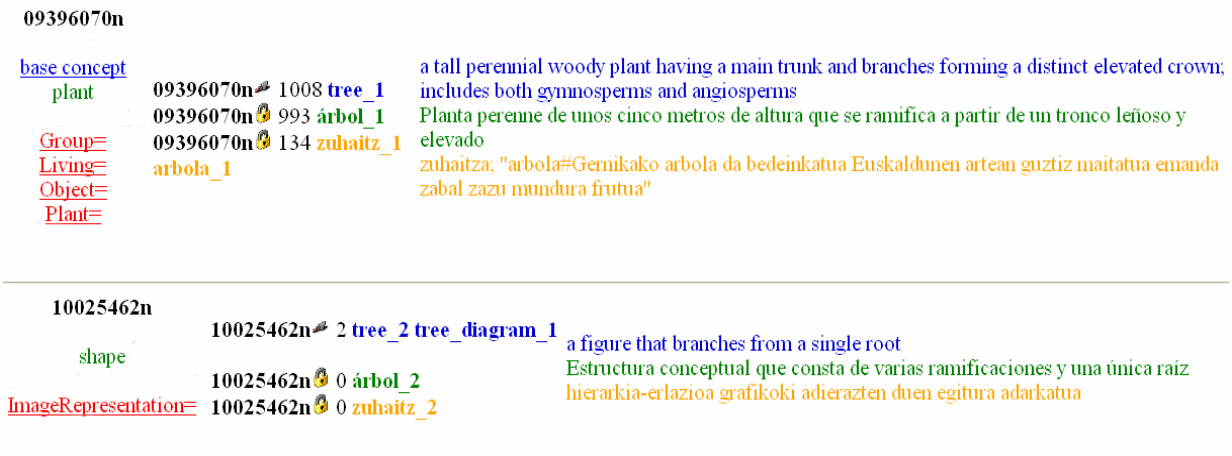
EuskalWordNet orain dela bost urte garatzen hasi ginen, eta etengabe aberasten doan EBLSa da. Gaur egun, EuskalWordNet-ek[12] 31.585 *synset* ditu (27.880 izen, 3.592 aditz eta 113 adjektibo). Bestalde, une honetan, izenen eskuzko orrazketarekin amaitzen ari gara eta aditzen orrazketarekin hasiberriak gara. Beraz, helburua da pixkanaka hiztegi osoari eta kategoria gramatikal desberdinei estaldura hedatzea. Gainera, ez da baztertzen, aurrerago, EuskalWordNet eta *Euskal Hiztegia*ren[13] arteko mapaketa edota bateratze bat egitea, hau da, EuskalWordNet *Euskal Hiztegia*tik eratortzen diren erlazio lexiko-semantikoekin aberastea.

EuskalWordNet-en erabilerak era askotakoak izan dira. Alde batetik, hiztegi eta thesaurus gisa erabili izan da. Hiztegi tradizionaletan bezala, EuskalWordNet-ek *synset* bakoitzeko glosa bat du, gehienetan adibide eta guzti[14]. Gainera, *synset* bakoitzean ale lexikal bat baino gehiago egon daitezkeenez, thesaurus bezala balia daiteke, adiera berdina adierazteko sinonimo desberdinak ditugulako.

Honenbestez, LNPri begira, EuskalWordNet-ek erabilera ugari izan ditu. Bakar batzuk aipatzeagatik, adieraren desanbiguazioan EuskalWordNet adieran oinarritutako ontologia denez, desanbiguazioan asko lagun dezake. Bestalde, adierak hierarkikoki antolatuta egoteak desanbiguazioaren atazan lagundu egiten du. Arlo honetan esperimentu ugari egin dira (Martínez, 2005).

[12] EuskalWordNet 1.6 bertsioaz ari gara.
[13] Sarasola, 1996.
[14] Une honetan, euskarazko izenen *synset*-en glosak garatzen ari gara.

10. Irudia: Kontsultarako bete beharreko eremuak.



11. Irudia: Kontsultaren emaitza.

Bestalde, EuskalWordNet-ekin etiketatutako corpusa oso lagungarria gerta daiteke, ordenagailuak corpusetik informazioa ikasteko, eta aldi berean, EBLSa aberasteko corpuseko informazioarekin. Honen froga ingelesez etiketatutako corpusa dugu: SemCor (Miller et al., 1994; Fellbaum et al., 2001). Hemendik abiatuta, egun, IXA taldean euskarazko corpus bat semantikoki etiketatzean ari gara: EuSemcor (Agirre et al., 2006).

3.5. EuskalWordNet: kontsulta

EuskalWordNet 1.6 bertsioa kontsultagarri dago http://ixa2.si.ehu.es/mcr/wei.html web orrian, eta interfazeren itxura 9. irudikoa da. Interfazearen erabileraren berri eman baino lehen, oinarrizko terminologia azalduko dugu.

3.5.1. Oinarrizko terminologia

Synset-a: Kontsultatu dugun hitzaren adiera ezberdin bakoitzari *synset* bat dagokio, eta interfazean marra batez bereiztuta agertzen da. 9. irudian ikus daitekeen bezala, *zuhaitz* hitzak bi *synset* ditu, hau da, bi adiera: 'arbola' eta 'diagrama'. Bestalde, s*ynset* bakoitzak bere zenbakia du (kasu honetan, 09396070n eta 10025462n). Hortaz, zenbaki hauen bidez kontzeptu zehatz horiek bakarrik adieraz daitezke.

Variant-a: *Synset* bakoitzean hizkuntza bakoitzeko dagoen ordaina da. Ordain bakoitzak adiera-zenbaki bat du. 9. irudian, adibidez, lehenengo *synset*-ean, *variant*-ak hurrengoak dira: ingelesezkoa, *tree_1*, gaztelaniazkoa *árbol_1* eta euskarazkoak *zuhaitz_1* eta *arbola_1*. Horrela bada, [ordaina+adiera-zenbakia] multzo horrek *variant*-a osatzen du, eta honen bitartez, kontzeptu zehatz bakarra adieraz daiteke.

3.5.2. EuskalWordNet: interfazea

Interfazea bitan banatua dago. Goiko aldean, egin beharreko kontsulta zehazteko baliagarriak diren eremuak ditugu; eta beheko aldean, kontsultaren emaitza.

Kontsulta bat egiterakoan, lehenengo testu-kutxatilan kontsultatu nahi den hitza, *synset*-a edo *variant*-a idatzi behar da (ikus 10. irudian A hizkia), eta ondoren, testu-kutxatilan idatzitakoa hitza, *synset* edo *variant*-a den zehaztu behar da testu-kutxatilaren azpiko eremuan (B hizkiaz 10. irudian). Esate baterako, kasu honetan, *zuhaitz* ordainaren *synset*-ak kontsultatu ditugu. Horretarako, *zuhaitz* hitza idatzi dugu testu-kutxatilan, eta ondoren, testu-kutxatilan idatzitakoa hitz bat (interfazean *word*) dela zehaztu dugu. Honekin batera, idatzitakoaren kategoria eta hizkuntza definitu behar dira. Gure adibidean, *zuhaitz* euskarazko izen bat denez, interfazean *noun* eta *BasqueWN1.6* aukeratu ditugu (B hizkiaz 10. irudian).

Ondoren, hitz horretaz zer jakin nahi dugun zehaztu behar da: sinonimoak, hiperonimoak, hiponimoak, meronimoak eta abar. Kasu honetan *zuhaitz* hitzak zer

synset dituen jakin nahi dugunez, *synonyms* erlazioa aukeratuko dugu (C hizkiaz 10. irudian).

Eta azkenik, hainbat kontrol-laukiei eraginda (D hizkiaz 10. irudian) kontsultaren emaitza pantailan informazio gehiago edo gutxiagorekin ikusteko aukera ematen zaigu: glosak ikustea ala ez (*Gloss*), *synset*-ak izan ditzakeen harreman semantiko mota guztiak ikustea ala ez (*Rels*), kontsultaren emaitza zer hizkuntzetan ikusi nahi den, eta abar[15].

Kontsulta honen emaitza 11. irudian ikus dezakegu. Alde batetik, *zuhaitz* hitzak bi *synset* dituela adierazten da. Lehenengo *synset*-a 'landare' adierari dagokio, eta bigarrena, berriz, 'diagrama' adierari.

Interfazearen ezkerraldera *synset* bakoitzeko informazio semantiko gehiago zehazten da:

Oinarrizko kontzeptuak (Base Concept): *Synset* batek marka hau badarama, hizkuntza guztietan dagoen oinarrizko kontzeptu bat dela adierazten da. *Zuhaitz* izenaren lehenengo *synset*-ak ('landare' adiera duena, alegia) marka hau darama.

Eremu Semantikoa (Semantic domain): *Synset*-aren eremu semantikoa zehaztu eta kontzeptuari buruzko informazioa osatzen duena da. Marka hau beti berdez adierazita dator. 11. irudian, *zuhaitz* ordainaren lehenengo *synset*-ak *plant* eremu semantikoa du, eta bigarrenak aldiz, *shape*.

Goi-ontologia (Top Ontology): Eremu semantikoa baino banaketa semantiko aberatsagoa da, WordNet ezberdinen goi aldeko *synset*-ak ezaugarri semantikoen arabera sailkatzea ahalbideratzen duena. Nolabait esateko, eremu semantikoen papera jokatzen du, nahiz eta motibazio linguistiko sakonagoak hartu diren kontuan. Marka hau beti gorriz adierazita dator. 11. irudian, *zuhaitz* hitzaren lehenengo *synset*-ak *Group, Living, Object* eta *Plant* ezaugarri semantikoak ditu, hau da, ezaugarri hauei esker jakin dezakegu *zuhaitz* lehenengo *synset*-ean talde bat osatzen duen eta bizirik dagoen gauza bat dela, eta gainera, landare mota bat dela.

Interfazearen erdialdean, kontsultarako aukeratutako hizkuntzen *variant* multzoa dago. Multzo honetan ere, bestelako informazioa jaso dezakegu. Esate baterako, *variant*-aren aurrean dagoen zenbakiak, *synset*-ak dituen hiponimo kopurua adierazten du. Hala, 11. irudiko *zuhaitz_1 variant*-a daraman *synset*-aren azpian, euskarazko 134 hiponimo daude.

Azkenik, interfazearen eskuinaldean, *synset*-aren hizkuntza bakoitzaren glosa dator (askotan ingelesekoa bakarrik dago, beste hizkuntzetan glosak ez dituztelako guztiz landuta).

[15] Web orrian (http://ixa2.si.ehu.es/mcr/wei.html) interfazeari buruzko argibide gehiago duen eskuliburua dago eskuragarri.

4. Ondorioak eta etorkizunerako lanak

Lan honetan EuskalWordNet ezagutza-base lexiko-semantikoa aurkeztu dugu, zuzenean helbide honetan atzitu daitekeena: http://ixa2.si.ehu.es/mcr/wei.html. EuskalWordNet garapenean dago oraindik. Izenak nahiko landuta daude, eta azken orrazketa bat egiten ari gara horien gainean. Aditzei dagokionez, kontzeptu garrantzitsuenak daude landuta bakarrik, eta orain ari gara aztertzen nola egin lanketa masiboa. Adjektiboei dagokionez oso gutxi landu ditugu eta aurreagorako utzi ditugu.

Gure azken helburua interpretazio semantikoa egitea da, eta hori aplikazioetan integratzea. Horretarako EuskalWordNet garatzen ari garen beste baliabideekin integratu behar da, beste hizkuntzetarako egiten ari den lez. Adiera desanbiguazioari aurre egiteko, adibidez EuskalWordNet-eko adierekin etiketatutako corpusa behar da, eta hori da hain zuzen ere EuSemcor corpusean egiten ari garena. Adierez gain, interpretazio semantikoan rol semantikoak desanbiguatu eta esleitu beharko dira, eta horretarako rol semantikoz etiketatutako corpus bat garatzen ari gara, EusPropBank deritzona (Civit et al., 2005).

Bestalde, interpretazio semantikorako beharrezkoak diren hainbat eta hainbat erlazio lexiko-semantiko (adibidez garagardoa garagarrez egina dagoela, okinaren lanbidea ogia egitea dela, arrantza pertsonek –eta astoek– egiten dutela, etab.). Aberasketa horretarako hiztegi eta corpusetatik ezagutza hori erdiautomatikoki erauzten saiatzen diren teknikak ere lantzen ari gara.

Azkenik, euskararen tratamendu morfosintaktikoa eta semantikoa lotu ahal izateko, Euskararen Datu Base Lexikalarekin integratzeko asmoa ere badago.

5. Aipamenak

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

Aduriz, I.; Alegria, I.; Arriola, J., Artola, X.; Díaz de Ilarraza, A.; Ezeiza, N.; Urkia, M. (1994). Euslem: un lematizador/etiquetador de textos en euskera. *Actas del X congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN).*

Agirre, E.; Ansa, O.; Arregi, X.; Arriola, J.; Díaz de Ilarraza, A; Pociello, E.; Uria, L. (2002). Methodological issues in the building of the Basque WordNet: quantitative and qualitative analysis. *Proceedings of First International WordNet Conference.*

Agirre, E.; Aldezabal, I.; Etxeberria, J.; Izagirre, E.; Mendizabal, K.; Pociello, E.; Quintian, M. (2006). Improving the Basque WordNet by corpus

annotation. *Proceedings of Third International WordNet Conference*.

Aldezabal, I. (2004). *Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa. Levin-en (1993) lana oinarri hartuta eta metodo informatikoak baliatuz*. UPV-EHUko tesi-lana.

Bresnan, J.; Kaplan, R.M. (1982). *The Mental Representation of Grammatical Relations*. Cambridge, Massachussetts: MIT Press.

Civit M., Aldezabal I., Pociello E., Taulé M., Aparicio J., Màrquez L. 2005 3LB-LEX: léxico verbal con frames sintáctico-semánticos. In *XXI Congreso de la SEPLN*. Granada (Spain).

Dowty, D. (1979). *Word Meaning and Montague Grammar*. Dordrecht: Reidel.

Fellbaum, C. (1998). *WordNet. An Electronic Lexical Database*. Cambridge, Massachussetts: MIT Press.

Fellbaum, C.; Palmer, M; Dang, H.T; Delfs, L.; Wolf, S. (2001). Manual and automatic semantic annotation with WordNet. *Proceedongs of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*.

Fernández, A.; Saint-Dizier, P.; Vázquez, G.; Kamel, M.; Benamara, F. (2002). The Volem Project: a framework for the construction of advanced multilingual lexicons. *Proceedings of Language Engineering Conference (LEC'02)*.

Fillmore, C.J; Baker, C.F. (2001). Framenet: Frame semantics meets the corpus. *Proceedings of WordNet and Other Lexical Resources Workshop.*

Jackendoff, R.S. (1990). *Semantic Structure*. Cambridge, Massachussetts: MIT Press.

Kipper, K.; Dang, H.T., Palmer, M. (2000). Class-based construction of a verb lexicon. *AAAI/IAAI* 691-696.

Levin, B. (1993). *English Verb Classes and Alternations. A Preliminary Investigation.* Chicago & London: The University of Chicago Press.

Martínez, D. (2005). *Supervised Word Sense Disambiguation: Facing Current Challenges*. UPV-EHU tesi-lana.

Miller, G.A. (1985). Wordnet: a dictionary browser. *Proceedings of the First International Conference on Information in Data*.

Morris, M. (1998). *Morris Hiztegia*.

Pociello, E. (2004). *Sintaxi-semantika elkargunea zenbait teoriatan: euskararen ezagutza-base lexiko-semantikorantz***.** UPV-EHUko ikerkuntza-lana.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, Massachussetts: MIT Press.

Sarasola, I. (1996). *Euskal Hiztegia*.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. *Language Typology and Syntactic Description, volume 3.* Cambridge University Press.

Vossen, P. (1998). EuroWordNet: A multilingual database with lexical semantic networks. Kluwer Academic Publishers.

