Euskararako ezagutza-base lexiko-semantikoaren eredu-hautaketa eta garapena: EuskalWordNet

Eneko Agirre, Izaskun Aldezabal eta Eli Pociello

IXA taldea, Euskal Herriko Unibertsitatea

http://ixa.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 counterparts EuroWordNet and the Multilingual Central Repository (MCR). This paper reviews the theoretical and practical aspects of the EuskalWordNet LSKB, as well as the steps followed in its construction.

Keywords: Natural Language Processing, lexical semantics, lexical knowledge bases, WordNet.

1 SARRERA ETA AURKEZPEN OROKORRA

Euskal Herriko Unibertsitateko Informatika Fakultateko IXA taldeak hamar urte baino gehiago daramatza Lengoaia Naturalaren Prozesamenduan (aurrerantzean LNP) lanean. Arlo zabal horren barruan, euskararen gaineko ikerketa aplikatua da gure xede nagusia, eta helburu horrekin orain arte, batez ere, morfologia (Agirre *et al.*, 1992) eta sintaxia (Aduriz *et al.*, 1998) 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:

- Euskal aditzen azpikategorizazioaren azterketa, hiztegi elebakar batean (Arriola, 2000) edo corpusetan oinarrituta (Aldezabal *et al.*, 2001).
- Euskal aditzen alternantzien eta klase semantikoen azterketa (Aldezabal, 2004).
- Aditzen hautapen-murriztapenen eskurapen automatikoa WordNet-en eta corpusetan oinarrituta (Agirre and Martínez, 2001).
- Adieraren desanbiguazioa (Martínez, 2005).

- Erlazio lexiko-semantikoen gauzatze sintaktikoa (Lersundi, 2005).
- Ezagutza lexiko-semantikoaren erabilera informazio bilaketan (Ansa *et al.*, 2005).

Lan hauei guztiei etekin handiagoa aterako litzaieke erabilitako baliabide eta deskribapen linguistiko guztiak ezagutza-base lexiko-semantiko (aurrerantzean EBLS) komun 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.

Behar horri erantzuteko, balizko EBLS horren hezurdura garatu eta definitzeari ekin genion: EuskalWordNet.

Lan honen helburuak euskararako EBLSa (EuskalWordNet) aurkeztea eta hau lortzeko jarraitutako prozedura deskribatzea da: euskararako baliagarria izan zitekeen eredu baten bila hasi ginenetik, aukeratutako eredua euskararako garatu eta aplikatu dugun arte.

Honenbestez, artikuluaren egitura hurrengoa da. 2. atalean EBLSen azalpen labur bat egiteaz gain, euskararako erabilgarria izan zitekeen EBLS baten proposamena egiteko zer eredu eta irizpidetan oinarritu garen azalduko dugu. Ondoren, 3. atalean euskarako EBLSa egiteko aukeratutako eredua aurkeztuko dugu (WordNet eta honen ildotik abiatutakoak) eta 4. atalean eredu hori aukeratu izanaren arrazoiak azalduko ditugu. 5. atalean EuskalWordNet-en inguruan egindako lana laburbilduko dugu. Eta azkenik, 6. atalean WordNet eredu gisa aukeratu izanaren ondorio batzuk aipatuko ditugu.

2 EBLSak

Lengoaia naturalen prozesamendu sintaktiko eta semantikoa egin ahal izateko, lexikoiak hitz-zerrenda izatetik EBLS izatera pasatu dira, hitzak adierari eta sintaxiari buruzko informazioaz hornituz. EBLS batean, hizkuntza ulertu ahal izateko, hitzei buruz jakin beharreko guztia egon beharko litzateke.

EBLSak baliabide lexikal egituratuak ditugu; hitzei dagokien informazioa elkarren arteko harreman-erlazioekin aberastuta dago. 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.

2.1 EBLSa definitzeko zailtasunak

EBLS bat lantzeko orduan zenbait zailtasunekin topatu gara. 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 and Kaplan, 1982; Fillmore and Baker, 2001; Miller, 1985; Kipper *et al.*, 2000, eta abar), eta askotan fenomeno linguistiko zehatz batera mugatuak daude.

Bestetik, definitzen zailak diren fenomeno linguistikoak zehaztu behar dira ale lexikal bakoitzeko, esaterako, alderdi semantikoa. Hala ere, egun oraindik iritzi ezberdinak daude ale lexikalen izaera semantikoa definitzerakoan: ale lexikalak berezko semantika du ala testuinguru sintaktikoaren eraginaren ondorioz jasotzen 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.

2.2 Euskararako EBLSaren aukeraketa eta ezaugarrien zehazpena

Gorago ikusi dugun bezala (2.1 atalean), EBLSen eraikuntzarako ez dago eredu bakarra, ez hizkuntzalaritza teorikoan ezta konputazionalean ere. Proposamen ugari daude, eta hizkuntzalaritza konputazionalaren kasuan, proposamen hauek arloetan zehar sakabanatuak daude.

Ondorioz, euskararako EBLS bat egiten hasi baino lehen, zenbait eredu edo formalismo aztertu ditugu. LNPren arloan jorratuak izan direnak interesatu zaizkigu bereziki —FrameNet (Fillmore and Baker, 2001), WordNet (Miller, 1985; Fellbaum, 1998), EuroWordNet (Vossen, 1997), MCR (Atserias *et al.*, 2004), 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).

Bestalde, euskarako EBLSaren diseinua irizpide batzuetara mugatu dugu, hau da, EBLS ereduak ondorengo baldintzak betetzea nahi dugu:

a) Hizkuntza bere osotasunean adierazten duen EBLSa izan behar du, ale lexikal bakoitzari dagokion adiera, klase semantikoa eta informazio sintaktiko-semantikoa (rol tematikoak, azpikategorizazioa, hautapen-murriztapenak, funtzio gramatikalak, kategoriak...) zehaztuta dituen EBLSa. Hizkuntzalaritza konputazionalaren ikuspegitik, zenbat eta lexiko aberatsagoa izan, orduan eta emaitza hobeak lortzen dira ataza konputazionaletan.

b) Ahal dela, teoria edo ikerlan bakar bati lotua ez dagoen EBLS eredua izatea, hau da, beste eredu edo formalismo batzuetatik edateko gaitasuna izatea. Behin eta berriz aipatu dugunez, EBLSaren eraikuntzarako ez dago eredu bat bakarra, ez hizkuntzalaritza teorikoan ezta konputazionalean ere, eta eredu bakarra jarraitzen duen EBLSra mugatzea arriskutsua izan daiteke. Beraz, ahalik eta *deskribatzaileena* den eredua interesatzen zaigu.

c) Konputazionalki inplementa daitekeen EBLSa izatea, hots, LNPn erabilgarria dena.

d) Aukeratutako eredu horretatik gertu beste lan konputazionalak egotea, gure EBLSa horien informazioarekin ere aberastu ahal izateko.

Gorago aipatutako ereduak irizpide hauen arabera aztertu eta ebaluatu ditugu, eta laburbilduz, azterketa honetatik ateratako ondorio nagusiak hurrengoak dira[1]:

[1] Azterketa zehatza ikusteko, jo bedi Pociello (2004*a*) lanera.

- **Jackendoff (1990):** Jackendoffen *Egitura Lexikal-Kontzeptualak* (ELKak) ale lexikalari buruzko informazio ugari dakar, batez ere sintaxi-semantika elkarguneari buruzko informazioa. Aditzak eta preposizioak azpikategorizazio aberatsena duten kategoriak direnez, eta hortik informazio sintaktiko-semantiko ugari lor daitekeenez, Jackendoff kategoria hauen adierazpenetara mugatu da, izenak, adjektibo eta adberbioak alde batera utziaz. Hortaz, esan dezakegu, eredu honek ez duela hizkuntza bere osotasunean adierazten. Bestalde, Jackendoffen lana, hizkuntzalaritza teorikotik aztertutako beste lanak bezala (Levin, 1993; Pustejovsky, 1995), ordura arte ez zegoen formalismo berri baten adierazle dira, beraz, ez daude beste formalismoetatik gertu, bakarrak dira. Hala ere, esan beharra dago, gerora Jackendoffen lan honetatik abiatuta egin direla lexiko batzuk. Esaterako, Dorr (1993, 1997) Jackendoffen lanean oinarritutako aditzen EBLS bat sortu zuen eta berarekin tutore sistemak eta itzulpengintza automatikoa landu zituen.

- **Levin (1993):** Levinek bere liburuan ingeleseko aditzen sintaxia eta semantika sakonki aztertzen du eta liburuan bertan landutako aditzen zerrenda ematen du, bakoitzaren adiera, klase semantiko, diatesi-alternantzia eta informazio sintaktiko-semantikoaren idatzizko deskribapenarekin batera. Euskararako EBLSaren proposamenerako baztertu egin dugu, batetik, aditzen deskribapen partziala bakarrik egiten duelako. Etorkizunean, informazio honetaz EBLSa aberastea aurreikusten dugu.

- **Pustejovsky (1995):** Autore honen ustez lexikoak sistema kognitiboaren oinarrizko egitura kontzeptuala adierazteko gai izan behar du, eta gainera, ahalmen sortzailea ere izan behar du. Horretarako, kategoria guztien deskonposaketan funtsatutako teoria proposatzen du. Alde horretatik beraz, euskararako nahi dugun EBLSarekin bat dator, hots, hizkuntza bere osotasunean aztertu nahi duen lana da. Hala ere, inplementazioari dagokionez, Pustejovskyren lanean oinarritutako ikerlan konputazional gutxi ezagutzen dugu, eta bere teoriatik abiatuta lexiko errealik ez dago. Hortaz, euskarako EBLSarentzat nahiko genukeen beste ezaugarrietako bat ez du.

- **LFG, GPSG eta HPSG:** Hizkuntzalaritza teoriko eta konputazionalaren erdibidean dauden lan hauen inguruan, hasteko, hizkuntzalaritza teorikoko lanei buruz esandako gauza bera errepikatuko dugu: hiru lan hauek ordura arte ez zegoen formalismo berri baten adierazle dira, beraz, ez daude beste formalismoetatik gertu, bakarrak dira. LFG, GPSG eta HPSG euskararekin erabiltzeko saiakera bat egin da (Gojenola, 1998), eta horretan agerian geratu zen hauen hasierako ingeleserako formulazioko ezaugarri eta erregela ugari ez zetozela bat euskararen izaera linguistikoarekin. Hortaz, LFG, GPSG eta HPSG euskarari aplikatu ahal izateko lan linguistiko handia egin beharko litzateke.

- **FrameNet**: FrameNet proiektuan (Fillmore and Baker, 2001) ingeleserako baliabide lexikografikoa eraikitzen ari dira, *Frame Semantics* (Fillmore, 1985) teorian oinarritua eta corpus errealeko datuekin lagunduta. Formalismo hau nahiz eta teoria bati lotua egon, corpus errealeko datuetan oinarritzen da, beraz, inplementa daitekeen EBLSa da, hots, praktikotasunera jotzen dute. EBLSa sortu eta lantzearekin batera, corpus etiketatu bat eratzen ari dira eta horrek hainbat erabilerari bidea zabaltzen die (baita konputazionalei ere). Horretaz gain, FrameNet EBLS publikoa da[2].

[2]FrameNet: www.icsi.berkeley.edu/framenet web.

Oso EBLS interesgarria da, batez ere ikuspegi konputazionaletik, LNPren arlo ezberdinen ikasketarako oso baliagarria delako. Baina eremu batzuetara (komunikazioa, legedia, hezkuntza...) mugatutako lexikoa da, denborarekin hizkuntza bere osotasunean adierazteko helburua duena. Beraz, epe luzerako EBLS da. Gure euskarako EBLSa, ez dugu eremu zehatzetan bakarrik mugatu nahi; hortaz, FrameNet ez da oraindik egokia osotasunari begira.

- **Volem:** Volem proiektuan (Fernández *et al.*, 2002) ingelesa, katalana eta frantsesa informazio sintaktikoz, argumentuz, azpikategorizazioz, rol tematikoez, hautapen-murriztapenez eta alternantziez definitzen dituzte. Eta adierak, batez ere, hiztegietan oinarriturik definitzen dituzte. Informazio hau guztia adierazteko eta antolatzeko Jackendoffen ELKak (1990) erabiltzen dituzte: alde batetik aditz eta preposizioen informazio sintaktiko-semantikoa adierazteko egokiak direlako, eta bestetik, ELKak eleaniztasunarekin bateragarria dela frogatu delako. Hala ere, nahiz eta EBLS hau Jackendoffen lanari alderdi semantikoa eta beste ikuspuntu teorikoak gehitu, aditz eta preposizioetara murrizten da, eta ondorioz, honek ere ez du hizkuntza bere osotasunean adierazten.

Azterketa horren ondorio gisa, eta aipatu ditugun irizpideen arabera, IXA taldearen beharretara gehiago egokitzen den EBLS formalismoa WordNet eta honen ildotik abiatuta garatu diren EuroWordNet eta MCR direla erabaki dugu. Hautaketa hau arrazoitu ahal izateko, lehenengo, 3. atalean, eredu hauen ezaugarriak laburki azalduko ditugu, eta ondoren, 4. atalean, hautaketaren arrazoiak emango ditugu.

3 AUKERATUTAKO EBLS EREDUA: WORDNET ETA HONEN ILDOTIK ABIATUTAKOAK

WordNet (Miller, 1985; Fellbaum, 1998) teoria psikolinguistikoetan oinarritua dagoen ingeleseko ezagutza-base lexikala da. Princeton-eko Unibertsitatean garatzen ari da — Cognitive Science Laboratory delakoan— George A. Milleren ardurapean.

Ingeleseko izen, aditz, adjektibo eta adberbioak *synonym set* edo ***synset***-etan (sinonimo multzotan) antolatuak daude, hauetako bakoitza kontzeptu lexikal bati dagokiolarik. Horrela, guk *synset*-a *adiera* gisa joko dugu. Esaterako, ingeleseko *tree* izenak WordNet-en bi *synset* (adiera) ditu:

(1) The noun "tree" has 2 senses:
1. tree (a tall perennial woody plant having a main trunk and branches...)
2. tree, tree diagram (a figure that branches from a single root; "genealogical tree")

Lehenengoa 'landare' (*plant*) *synset*-ari dagokio eta bigarrena, berriz, 'diagrama' (*diagram*) *synset*-ari. Lehenengo *synset*-a ale lexikal bakar batez osatua dago (*tree*), hots, *tree* izenak *synset* horretan ez du beste sinonimorik. Bigarrenak, ordea, *synset*-ean *tree* ale lexikalaz gain, beste ale bat ere badu (*tree diagram*), horrela, bi ale lexikal horiek (*tree* eta *tree diagram*) sinonimoak dira.

Ildo honetatik, WordNet-eko erlazio semantiko garrantzitsu bat **sinonimia** da; ezagutza-basearen oinarria ale lexikalaren adieran baitago, eta adiera hori ale lexikal batek baino gehiago duenean, ale lexikalak multzokatu egiten dituztelako.

WordNet-eko sinonimiaz hitz egiterakoan, kontuan izan behar da ez dela gauza bera sinonimia eta hitzak bata besteaz elkar trukatzea. Hau da, WordNet-eko *synset*-a osatzen duten sinonimoak beraien artean truka daitezke, baina **testuinguru batzuetan** bakarrik.

> The more modest claim is that WordNet synonyms can be interchanged in some contexts. To be careful, therefore, one should speak of synonymy relative to a context. (Fellbaum, 1998, 24. or)

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

Hiperonimia-hiponimia erlazioak *synset* orokorrenak *synset* zehatzagoekin lotzen ditu[3]. (2) eta (3) adibideetan (1)en hiperonimoak eta hiponimoak ikus ditzakegu hurrenez hurren[4]:

(2) Sense 1
tree (a tall perennial woody plant having a main trunk and branches...)
=> woody plant, ligneous plant – (a plant having hard lignified tissues...)
=> vascular plant, tracheophyte – (green plant having a vascular system...)
=> plant, flora, plant life – (a living organism lacking the power of locomotion)
=> life form, organism, being, living thing – (any living entity)
=> entity, something – (anything having existence (living or nonliving))

Sense 2
tree , tree diagram (a figure that branches from a single root; "genealogical tree")
=> plane figure, two-dimensional figure (a 2-dimensional shape)
=> figure (a combination of points and lines and planes that form a visible palpable shape)
=> shape, form (the spatial arrangement of something as distinct from its substance)
=> attribute (an abstraction belonging to or characteristic of an entity)
=> abstraction (a general concept formed by extracting common features...)

(3) Sense 1
tree (a tall perennial woody plant having a main trunk and branches...)
=> yellowwood, yellowwood tree (any of various trees having yellowish wood...)
=> lancewood, lancewood tree, Oxandra lanceolata (source of most of the lancewood of commerce)
=> Guinea pepper, negro pepper, Xylopia aethiopica (tropical west African evergreen tree...)
=> anise tree (any of several evergreen shrubs and small trees of the genus Illicium)
=> winter's bark, winter's bark tree, Drimys winteri (South American evergreen tree...)
=> zebrawood, zebrawood tree (any of various trees or shrubs having mottled or striped wood)
=> granadilla tree, granadillo, Brya ebenus (West Indian tree yielding a fine grade of green ebony)
=> acacia (any of various spiny trees or shrubs of the genus Acacia)
=> ...
Sense 2
tree, tree diagram (a figure that branches from a single root; "genealogical tree")
=> cladogram (a tree diagram used to illustrate phylogenetic relationships)

[3]Ingelesez *IS-A relation* bezala ere ezagutzen da, hots, *x is y*.

[4]Adierazpen guztiak WordNet 2.0 bertsiotik hartu ditugu (http://www.cogsci.princeton.edu/cgi-bin/webwn).

(2) adibidean *tree* izenaren hiperonimoak ditugu. Lehenengo *synset*-a ('landare') kontuan hartuz gero, *woody plant* mota bat bezala definitzen da; *woody plant vascular plant* mota bat bezala; *vascular plant*, aldi berean, *organism* mota bat bezala, eta azkenik, *organism entity* mota bat bezala. Ondorioz, *tree*, bere lehenengo *synset*-ean, *entity*, *organism*, *vascular plant*, eta *woody plant* bat izango da.

Tree-ren beste *synset*-aren ('diagrama') sailkapenarekin berdin-berdin gertatzen da, baina bere hiperonimoak 'diagrama' adierari lotuak egongo dira.

Hiponimoak hiperonimoen zehaztapenak dira. Hortaz, (3) adibidean, *tree* izenaren lehenengo adieraren zehaztapen gisa zuhaitz motak agertzen dira (*yellowood, acacia...*), eta bigarren adieran, aldiz, diagrama motak (kasu honetan bakarra, *cladogram*). Horrela bada, WordNet 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 erlazioa semantiko nagusi baten arabera antolatzen da. Izen eta aditzen kasuan erlazio semantiko nagusia hiperonimia-hiponimia da[5]. Adjektibo eta adberbioek, berriz, sinonimia-antonimia dute ardatz gisa beraien antolakuntzan. (4) adibidean, *properly* adberbioaren antonimoa ikus dezakegu (*improperly*):

(4) Sense 1
properly , decently, decent, in good order, right, the right way (in the right manner)
=> improperly (in an improper way)

WordNet-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, zuhaitz mota desberdinak *synset* baten azpian jasota daude (*tree* izenaren 'landare' *synset*-ean, alegia), hortaz, zuhaitzen motak jasotzen dituen klase semantikoa *synset* horren bitartez adieraz daiteke.

3.1 Bestelako erlazio semantikoak

Sinonimia eta hiperonimia-hiponimia/troponimia erlazio semantikoetaz gain, WordNet-ek beste asko landu ditu. Hemen batzuen aipamen laburra egingo dugu[6].

Izenak lotuak egon daitezke ondorengo erlazio semantikoen bidez:

- **Zati-osotasun harremana (*Part-whole relation*)**:

Zatia eta osotasuna harremanetan jartzen dituen erlazioak dira. Batetik, **meronimia** dago, *x is part of y* definizioari jarraitzen diona; *finger hand*-en zati bat da eta *hand* aldi berean *arm*-ena:

[5]Aditzen kasuan, hiperonimia-troponimia erlazioz hitz egiten da. Aditzek hiponimiaren ordez **troponimia** erabiltzen dute. Honen arrazoia da aditzak ezin direla *IS-A* motakoak izan, baizik eta *to x is to y in some particular manner* motakoak. Hortaz, aditz hiperonimo baten (*walk*) troponimoak aditz hiperonimoak adierazten duena egiteko moduak izango dira (*trot*, *march*...). Argibide gehiagorako jo Fellbaumen lanera (1998).

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

(5) 1 of 2 senses of "finger"
Sense 1
finger (any of the terminal members of the hand; "her fingers were long and thin")
PART OF: hand, manus, hook, mauler, mitt, paw (the extremity of the superior limb)
PART OF: arm (technically the part of the superior limb between the shoulder and...)

Eta bestetik, *x has a y (as a part)* definizioarekin bat badator, orduan, **holonimia** erlazioaren bidez lotzen dira; adibidez, *wheeled vehicle* batek *wheel*-ak dauzka:

(6) 2 of 6 senses of "wheel"
Sense 1
wheel (a simple machine consisting of a circular frame with...)
PART OF: wheeled vehicle (a vehicle that moves on wheels ...)

- **Antonimia**:

Izen batzuek antonimoak dituzte eta erlazio semantiko honek lotzen ditu:

(7) 1 sense of "victory"
Sense 1
victory, triumph (a successful ending of a struggle or contest)
=defeat, licking (an unsuccessful ending)

Aditzen hierarkian erlazio semantiko nabarmenetako bat *entailment* deritzona da (*V1 logically entails V2* edota *snore entails sleeping*).

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

(8) 1 of 6 senses of "nice"
Sense 1
nice (vs. nasty) (pleasant or pleasing or agreeable in nature or appearance; "nice manners")
=> niceness (the quality of nice)

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

3.2 Erabilera

WordNet-ek 115.424 *synset* ditu (79.689 izen, 13.508 aditz, 18.563 adjektibo eta 3.664 adberbio)[7].

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

Honenbestez, LNPri begira, WordNet-ek erabilera ugari izan ditu. WordNet-eko web orrian agertzen den bibliografian[8] hau erakusten duten 2.000 artikulu inguru daude. Guk arlo bakoitzetik garrantzitsuenak baino ez ditugu aipatuko:

[7] WordNet-en 2.0 bertsioaz ari gara.
[8] Ikus http://engr.smu.edu/ rada/wnb/ web orrian.

- **Adieraren desanbiguazioan:** WordNet adieran oinarritutako ontologia denez, desanbiguazioan asko lagun dezake. Bestalde, adierak hierarkikoki antolatuta egoteak desanbiguazioaren atazan lagundu egiten du. Arlo honetan esperimentu ugari egin dira (Miller *et al.*, 1994; Agirre and Martínez, 2000, besteak beste).

- **Itzulpen automatikoan** aipagarria da Dorrek (Dorr, 1997) Jackendoffen teorian eta WordNet-eko adieretan oinarrituta egindako lana. Honetaz gain, ikerlan gehiago ere egin dira, esate baterako, Rigau *et al.* (1995).

- **Informazio erauzketan:** WordNet lagungarria izan daiteke erabiltzaileak bilaketa bat egiterakoan, eta beharrezkoa zaion edukia duen dokumentua lortzeko. Bilaketan erabilitako hitzak dokumentuan daudenekin bat etorri behar dute, emaitza egokia lortzeko. WordNet-ek horretan lagun dezake. Zenbait saiakera egin dira, esaterako, Magnini and Strapparava (2001), Mandala *et al.* (1998), besteak beste.

  Bestalde, WordNet-ekin etiketatutako corpusa oso lagungarria gerta daiteke, ordenagailuak corpusetik informazioa ikasteko. Honen froga ingelesez etiketatutako corpusa dugu: SemCor (Miller *et al.*, 1994; Fellbaum *et al.*, 2001).

- **Galdera-erantzunean eta informazio-bilaketa:** *synset*-en arteko harremanek galdera bati dagozkion erantzunak ezagutzea laguntzen dute (Ansa *et al.*, 2005; Vossen *et al.*, 2006, eta abar).

WordNet abiapuntutzat hartuta, beste ezagutza-base batzuk egin dira: EuroWordNet (Vossen, 1998) eta *Multilingual Central Repository* (MCR)(Atserias *et al.*, 2004). Oinarri bera erabili arren, bere gainean aberasketa batzuk egin dituzte. Hurrengo ataletan (3.3 eta 3.4) oso laburki hauen berri emango dugu.

3.3 EuroWordNet

EuroWordNet proiektua (Vossen, 1998) 1996an hasi eta 1999 urteraino luzatuko den 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 (ikus 3. atala), hots, Princeton-en ingeleserako egindako WordNet-aren *synset*, harreman semantiko eta hierarkian oinarritu dira beraien WordNet-a sortzeko.

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.

WordNet-en egitura, harreman eta *synset*-etan oinarritu arren, WordNet-ek ez zituen ezaugarri batzuk EuroWordNet-en gaineratu dira. Aldaketarik aipagarrienak hurrengoak dira[9]:

[9] Argibide gehiago Vossenen lanean (1998).

- **Erlazio semantikoen aberasketa**:

WordNet-eko erlazio semantiko batzuk findu egin dituzte eta beste erlazio semantiko batzuk aberastu. EuroWordNet-ek hizkuntza barneko erlazio mota gehiago ditu, eta batez ere, morfologikoki aldatzen diren kategoria ezberdinen arteko erlazioak ugaritu dituzte (*nice* eta *niceness* bezalakoak, alegia).

Bestalde, EuroWordNet-ek ez du WordNet-en interfaze informatikoa; EuroWordNet-ena interfaze berria da, hizkuntza bakoitzeko WordNet-ak erlazio berriak gehitzeko aukera duelarik.

- **Hierarkiaren aberasketa:**

WordNet-ek ez zuen hierarkiari Domeinu-ontologia (*domain ontology*) eta Goi-ontologia bat (*Top ontology*) gehitu dizkiote.

Lehenbizikoak, *synset*-ak domeinuen arabera antolatzen ditu: *free time*, *restaurant*, *traffic*, eta abar. Esate baterako, *jokatu* aditzak kirola adierazten duenean (*futbolean jokatu* diogunean, adibidez), *synset* horrek *free time* domeinuaren marka eramango du; *zuzen jokatu* esan nahi dugunean, ordea, adiera horri dagokion *synset*-ak *psychology* marka izango du[10].

Bigarrenak, WordNet ezberdinen goi aldeko *synset*-ak oinarrizko ezaugarri semantikoen arabera sailkatzea ahalbidetzen du[11], eta nolabait esateko, EuroWordNet-eko domeinuen papera jokatzen du, nahiz eta motibazio linguistiko sakonagoak hartu diren kontuan. Hau da, tasun linguistikoak ([+/- bizidun], [+/- egile] adibidez) kontuan hartzen dituen ontologia da eta WordNet tasun hauen arabera eraikitzen da. Hortaz, ale lexikal bat [+biziduna] bada Goi-ontologiaren [+biziduna] adabegiaren azpian kokatuko da eta [-biziduna] bada, aldiz, [-biziduna] ezaugarriaren behean.

Oinarrizko ezaugarri semantikoak definitzerakoan, EuroWordNet-en sortzaileak hizkuntzalaritzan egon diren zenbait sailkapen semantikoen eredutan oinarritu dira: Dowty (1979), Levin (1993), Lyons (1977), eta Pustejovskyren ereduan (1995), besteak beste.

Guztira, 63 ezaugarri semantikok osatzen dute Goi-ontologia hau, eta Lyonsen lanari (1977) jarraituz hiru maila bereizi dituzte:

- **Lehenengo mailako entitateak (*1st Order Entity*)**: Zentzuen bidez antzeman daitekeen eta denboran/lekuan antzeman daitekeen entitateak dira (*animalia, objektu, substantzia* eta abar bezalako ale lexikalak).

- **Bigarren mailako entitateak (*2nd Order Entity*)**: Edozein egoera estatiko edo dinamiko, zentzuen bidez objektu fisiko bezala ezagutu ezin daitekeena. Denboran koka daitezke eta "gertatu" egiten dira "existitu" baino gehiago (*gertatu, hasi, jarraitu, izan, eduki*, amaitu bezalako ale lexikalak). Beraz, ekintzak, prozesuak eta egoerak adierazten dituzten ale lexikalak maila honen azpian egongo dira (*gertatu, hasi, jarraitu, izan, eduki, amaitu* bezalakoak).

- **Hirugarren mailako entitateak (*3rd Order Entity*)**: Ikus ezin daitezkeen proposizioak dira, denbora eta lekuan koka ezin daitezkeenak. Proposizioak direnez, egia edo gezur bezala uler daitezke, erreal edo irrealtzat baino (*ideia, pentsamendu, informazio, teoria, plana* bezalakoak).

[10] Domeinuen sailkapena ez da EuroWordNet-eko interfazean ikusten, beste fitxategi batzuetan daude.

[11] Goi-ontologiak goi aldeko *synset*-ak sailkatu arren, hauen azpian dauden *synset*-ak ere sailkapen hori mantentzen dute, beraien hiperonimoen ezaugarriak heredatzen dituztelako.

Goi-ontologiako maila hauen arteko desberdintasuna ageriagoa da hauek adierazteko erabiltzen diren kategoria sintaktikoei erreparatzen badiegu:

- **Lehenengo mailako entitateak (*1st Order Entity*)**: izen konkretuak
- **Bigarren mailako entitateak (*2nd Order Entity*)**: izenak, aditzak eta adjektiboak
- **Hirugarren mailako entitateak (*3rd Order Entity*)**: izen abstraktuak

1. irudian Goi-ontologiaren hierarkia guztia zerrendatua ikus dezakegu. Larbubilduz, esan daiteke EuroWordNet-ek WordNet-eko *synset* eta erlazio semantiko guztiak dituela, eta gainera, *synset* horiek Goi- eta Domeinu-ontologiekin aberastuta daudela.

3. atalean *tree* izena hartu dugu adibide gisa. EuroWordNet-en *tree* izenaren *synset* horiek berak egongo dira, baina horretaz gain, interfazean bertan Goi-ontologiako tasunak ere ikusten ditugu. 2. irudian *tree* izenaren *synset* bat dugu —'diagrama' adierari dagokiona, hain zuzen ere— bere hiperonimoekin. Irudi honetan *Image Representation* eta *Physical Property* dira Goi-ontologiako markak. Nahiz eta Goi-ontologiako tasunak *tree* izenaren *synset*-ean bertan ez egon, bere hiperonimoetatik herentziaz jaso egiten ditu. Hau da, *synset* horren *figure_6* hiperonimoak *Image Representation* tasuna du, eta aldi berean, *figure_6* horren hiperonimoak (*form_6*) *Physical Property* tasuna. Hortaz, *tree* izenaren *synset*-ak ezaugarri horiek guztiak herentziaz jasotzen ditu. Beste hitz batzuetan esanda, EuroWordNet-en tasun hauek ez dituzte *synset*-ez *synset* adierazten, defendatzen dutelako hierarkiari esker herentziaz jaso daitezkeela.

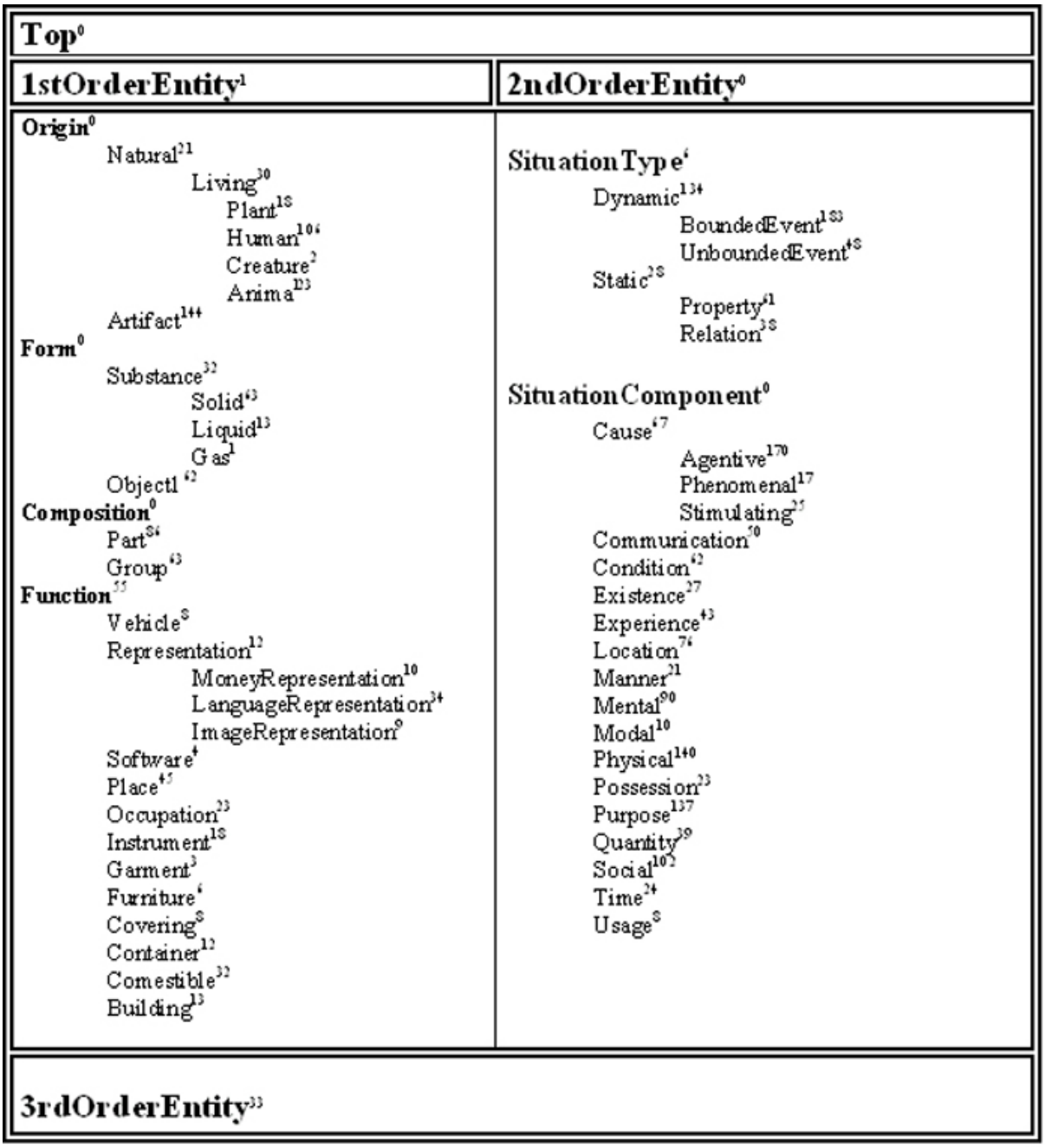
Azalduriko ezaugarriek —eleaniztasunak eta ikerkuntzarako erabilgarria izateak, alegia— oso egoki bihurtu dute ezagutza-base hau LNPren barnean erabiltzeko, batik bat, informazio erauzketa elebakar eta elebidunerako (Vossen, 1997). Eta arrazoi horregatik, gaur egun, hainbat WordNet berri sortzen ari dira (katalana, portugesa, grekoa, suediarra, errumaniarra, bulgariarra, norvegiarra, lituaniarra, errusiarra...) EuroWordNet-en ezagutza-basean oinarrituta. IXA taldean ere euskararako WordNet-a garatzen ari gara (Agirre *et al.*, 2002). EuroWordNet www.illc.uva.nl/EuroWordNet web orrian dago eskuragarri.

3.4 The Multilingual Central Repository (MCR)

The Multilingual Central Repository (MCR) interfaze eleanitza da, non Europa Batzordeko "MEANING: Developing Multilingual Web-Scale Language Technologies" (IST-2001-34460) proiektuan (Rigau *et al.*, 2003) aztertu den informazio guztia integratzen den. Ezagutza-base honek EuroWordNet-en eredua jarraitzen du.

MCRk bost hizkuntzetako WordNet-ekin egiten du lan: euskara, katalana, ingelesa, italiera eta gaztelania. MCR bost hizkuntza horien izen, aditz, adjektibo eta adberbioen adieren inbentarioa da, eta EuroWordNet-en eredua jarraitzen duenez, hizkuntza guztiak lotuta daude. Horregatik, hizkuntza bateko *synset* batekin beste hizkuntzetakoa ere ikusgarri dago.

EuroWordNet WordNet-en garapena denez, MCR EuroWordNet-en bertsio aurreratuagoa ere da. MCR WordNet eta EuroWordNet-en informazioaz baliatzen da, eta honetaz gain, informazio berria dakar:

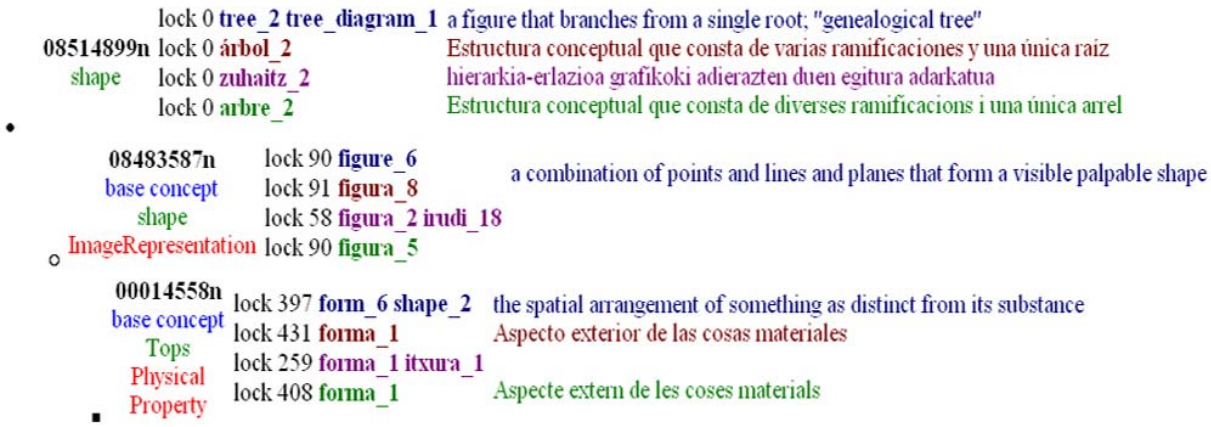


1. irudia: EuroWordNet-eko Goi-ontologia.

- **Domeinu-ontologiaren bertsio aberatsago bat:**

EuroWordNet-eko domeinuak ugaritu eta orraztu dituzte[12], hierarkian egon zitezkeen irregulartasunak gainditzeko. Bestalde, entitate edo izen bereziei domeinuak esleitu dizkiete,

[12]EuroWordNet-eko hainbat domeinu gehiago zehaztu dituzte, "azpidomeinuak" sortuaz. Esate baterako, *jokatu* aditzak kirol adiera duenean, EuroWordNet-eko *free time* domeinua, *sport* jarrita zehaztu dute.



2. irudia: *Tree* izenaren *synset* bat eta bere hiperonimoak EuroWordNet-eko interfazean.

eta horren ondorioz, egitasmo horren emaitza domeinuka antolatutako izen berezi eta entitateen ezagutza-base bat da.

- **Hautapen-murriztapenak:**

Hautapen-murriztapenen azterketa automatikoa egin da (Agirre *et al.*, 2003; Carroll *et al.*, 2003), hau da, teknika konputazional desberdinak erabiliaz zenbait corpusetatik (*Egunkaria*, *British National Corpus* eta *SemCor*-etik hain zuzen ere) aditzen hautapen-murriztapen usuenak zeintzuk diren ikasi dira, ondoren, MCRn erlazio semantiko gisa hauen berri izateko.

Hortaz, MCR ezagutza-basean aditzen hautapen-murriztapenak kontsulta daitezke *Role* erlazio semantikoa erabilita. Erlazio semantiko honek, zenbait aukera eskaintzen ditu, begiratu nahi dugun erlazioaren arabera. Esate baterako, aditz baten *egile* hautapen-murriztapena zein den ikusteko, *role agent* erlazioa erabil dezakegu; *gaia* ikusteko, aldiz, *role patient* erlazioa. Hauek dira dauden beste erlazioetako batzuk: *role direction*, *role instrument*, *role location*, *role source location*, *role target direction*...

Corpusetako datuetan oinarrituz, erlazio semantiko hauen guztien bitartez aditz bakoitzak ager daitezkeen ale lexikoak eta har ditzaketen rol tematikoak bereizteko gai dira. Ondorioz, MCRn aditzaren rol tematikoen berri ematen duen erlazio semantikoa dago.

MCRn, ale lexikalak ontologian kategoriaka antolatuta daudenez (WordNet eta EuroWordNet-en bezala) *Role* erlazioak inplizituki ere azpikategorizazioaren berri ere eman dezake. Hortaz, *Role* erlazioak *edan* aditzaren *Role patient* erlazioaren bidez aditz horri buruz esango liguke bere hautapen-murriztapena *edari* dela[13], izena kategoria duela eta bere rol tematikoa *gaia* (*patient*) dela. Hurrengo urratsetan, sintaxi-semantikari buruzko informazio gehiago, hala nola, funtzio gramatikalak, txertatzeko asmoa dago.

MCRren kontsultarako interfazea ondorengo web orrian dago: http://nipadio.lsi.upc.es/cgi-bin/public/wei1.consult.perl

Horrela bada, WordNet eta EuroWordNet ezagutza-basearen ildotik jarraituz, MCRk erakutsi du hasieran egitasmo semantiko eta psikolinguistiko soilekin burutu zen ezagutza-basea, informazio sintaktiko-semantikoa jasotzeko ere baliagarria izan daitekeela. Proiektu

[13] *Edan* aditzaren hautapen-murriztapena *edari* eta honen hiponimo guztiak ere badira.

honen hurrengo urratsetan MCR informazio sintaktiko-semantiko gehiagorekin (azpikategorizazioa, diatesi-alternantziak, Dorr-en ELKak (1997), erlazio semantiko konplexuagoak eta abar) osatzeko asmoa dago.

4 WORDNET ETA WORDNET-ETIK ABIATUTAKOAK AUKERATZEKO ARRAZOIAK

2.2. atalean euskararako EBSLarentzat nahi ditugun ezaugarriak zehaztu ditugu. Ondoren azalduko dugun bezala, WordNet eta WordNet-etik abiatutakoak (EuroWordNet eta MCR, alegia), ezaugarri hauekin bat datoz, eta arrazoi horregatik hain zuzen ere, aukeratu genituen guk garatu nahi dugun EBLSaren eredu gisa.

a) **Hizkuntzaren ikuspuntu orokorra:** WordNet (EuroWordNet eta MCR) lexiko zabala eta garatua duten EBLSak dira. Hauek adieran oinarritutako ontologiak dira, hizkuntzaren lexikoa ezagutza-base batean jaso nahi dutenak, ale lexikalak, ale lexikalen adierak, klase semantikoak, kategoriak, eta hauen guztien arteko erlazio semantikoak kontuan izanda (3, 3.3 eta 3.4 ataletan azaldu dugun bezala). Noski, hizkuntzaren lexikoak ez du mugarik eta etengabe garatzen dauden ezagutza-baseak dira. Hala ere, hizkuntzaren ikuspuntu orokorra eman dezaketen ezagutza-baseak ditugu. Esate baterako, WordNet-ek 115.424 *synset* ditu (79.689 izen, 13.508 aditz, 18.563 adjektibo eta 3.664 adberbio)[14]. MCRk WordNet ezagutza-basearen tamaina bera du, baina erlazio semantiko gehiagorekin (milioi bat inguru).

b) **Eredu deskribatzailea:** WordNet ez dago teoria bakar bati lotua, hots, teoria ezberdinek erabil dezaketen EBLSak dira. Bestalde, EuroWordNet eta MCR berak, WordNet-en garapenak dira, WordNet beste oinarri eta ikuspuntu teoriko eta konputazionaletatik informazio gehiagorekin aberastu dutelarik.

c) **Inplementazioa:** WordNet, EuroWordNet eta MCR inplementatutako EBLSak dira, hots, praktikotasunera jotzen dute. EBLSa sortu eta lantzeaz gain, hainbat erabilera izan ditzakeen ezagutza-base publikoak dira, kontsultagarriak, alegia. Hauek aztertzerakoan aipatu ditugu beraien erabilera nagusiak (hiztegi eta thesaurus gisa adibidez).

   EuroWordNet eta MCRrekin inplementatzeko aukerak areago doaz, EBLS hauek eleanitzak direlako, ingeleseko WordNet-ari beste hainbat hizkuntza gehitu dizkiotelako (daniera, italiera, gaztelania, alemana, frantsesa, txekiera, estoniera...) eta horien artean euskara txertatzen ari gara (Agirre *et al.*, 2002).

d) **LNPn erabilgarria:** Hiru EBLS hauek oso erabiliak izan dira LNPren arlo oso ezberdinetan: galdera-erantzunean, informazio erauzketan, itzulpen automatikoan eta adiera desanbiguazioan besteak beste (hauei buruzko argibide gehiago 3.2, 3.3 eta 3.4 ataletan).

e) **Beste lan batzuekiko gertutasuna:** Formalismo eta lan teoriko askok, gerora, WordNet eta EuroWordNet-en adiera edo/eta klase semantikoekin aberastu dituzte[15]. Esate baterako, Dorrek (1997) Jackendoffen lanarekin egin duen bezala. Dorrek

[14] WordNet-en 2.0 bertsioaz ari gara.

[15] MCR orain dela gutxiko EBLSa izanda, oraindik ez da horrela erabili.

Jackendoffen *Egitura Lexikal-Kontzeptualetan* (ELKetan) oinarritutako EBLSa eraiki du eta ELK hauek WordNet-eko adieretara lotuak daude. Lan horretan bertan, Dorrek Levinen aditz-klaseetako aditzak ere WordNet-eko aditzekin lotzen ditu. Ildo honetatik jarraitu duen formalismoa Volem izan da. Bestalde, Pustejovskyren lexikoaren ezaugarri batzuk WordNet-ekoekin lotzeko saiakera ere egin da (Buitelaar, 1998). Formalismo ezberdin hauen arteko uztardura oso baliagarria eta aberatsa da, beti ere, WordNet-en ildotik euskararako egingo den EBLSak hauez guztiez balia baitaiteke euskarako EBLSan informazio gehiago eransteko. Beraz, garbi dago WordNet eta EuroWordNet-ek LNPren arloan baliabide oso erabiliak izan direla, eta egun oraindik hainbat esperimentu eta ikerlanetarako iturburu direla.

4.1 Gabeziak

Hala eta guztiz ere, aukeratutako ildo honek gabezia batzuk badituela jabetzen gara. Esate baterako, WordNet eta EuroWordNet ezagutza-base hauetako informazio sintaktiko-semantikoa mugatua da. Adibidez, ez dituzte azpikategorizazioa, hautapen-murriztapenak eta rol tematikoak zehazten. Hala ere, hautapen-murriztapenen kasuan, WordNet eta EuroWordNet-en hauek inplizituki badaude, baina EBLSan oraindik ez daude erlazio baten bidez esplizituki adierazita, hau da, WordNet eta EuroWordNet-eko EBLSek ez dute eskaintzen *edan* eta *edari* sarrera lexikalak hautapen-murriztapen gisa lotzeko biderik. Hau oztopo bat da euskarako EBLS bat hauetan oinarrituta egiteko, lanaren hasieratik esan dugun bezala (2.2 atalean), euskarako EBLSan ale lexikalen adierez gain, hauen informazio sintaktiko-semantikoa adierazita etortzea nahiko genukeelako.

Dena den, MCRn horrelako informazioa esplizitua egiten saiatzen dira, esate baterako, hautapen-murriztapenak MCRn adierazita daude *Role* erlazio semantikoaren bitartez (*Role agent*, *Role patient* besteak beste[16]). Hortaz, erlazio semantiko honen bitartez ale lexikal batekin ager daitezkeen ale lexikal motak, hots, hautapen-murriztapenak, eta azken hauek har ditzaketen rol tematikoak bereizteko gai dira. MCRn ale lexikalak ontologian kategoriaka antolatuta daudenez (WordNet eta EuroWordNet-en bezala) *Role* erlazioak inplizituki ere azpikategorizazioaren berri ere eman dezake. Hortaz, *Role* erlazioak *edan* aditzaren *Role patient* erlazioaren bidez aditz horri buruz esango liguke, bere hautapen-murriztapena *edari* dela[17], honek izena kategoria duela, eta bere rol tematikoa *gaia* (*patient*) dela. Hurrengo urratsetan, sintaxi-semantikari buruzko informazio gehiago txertatu nahi da, hala nola, funtzio gramatikalak.

Euskarako EBLSa MCRren ereduan eraikita, honek WordNet eta EuroWordNet-en hezurdura izango du, hots, adieraka antolatutako EBLS semantiko eleanitz baten abantailak izango ditugu, eta gainera, bi ezagutza-base hauetan dagoen informazioarekin batera, MCRn gehitu (eta gerora gehituko) den informazio sintaktiko-semantikoa eskuragarri dugu.

5 EUSKALWORDNET

Hala, IXA taldea garatzen ari den euskararako EBLSa WordNet-en eta honen ildotik sortutakoetan oinarritua da: EuskalWordNet deritzoguna. Beraz, gure EBLSa *synset*-etan oinarritutako hierarkia kontzeptuala da (ikus 2. atala).

[16] Ikus 3.4 atala.

[17] *Edan* aditzaren hautapen-murriztapena *edari* eta honen hiponimo guztiak dira.

Atal honetan EuskalWordNet-en inguruan egindako lana laburbilduko dugu. Hasteko, EuskalWordNet garatzeko erabilitako metodologiak deskribatuko ditugu (5.1, 5.2 eta 5.3 atalak); ondoren, EuskalWordNet-en egungo kopuruak aipatuko ditugu (5.4), eta azkenik, EuskalWordNet-en interfazea eta bertan egin daitezkeen hainbat eragiketa aurkeztuko ditugu (5.5).

5.1 EWN eta MCRtik nola abiatu

Behin euskarako EBLS egiteko oinarrituko garen eredua erabakita, eta EBLS hori ingeleserako sortu dela ikusita (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, ikus 3.3. atala).

Bi hurbilpenen alde onak eta txarrak aztertu ondoren, euskararako WordNet-a egiteko EuroWordNet abiapuntutzat hartzea erabaki dugu (Agirre et al., 2002), hau da, EuroWordNet-eko ingelesezko kontzeptuei euskarazkoak lotuz, eta ez dauden euskarazko kontzeptuak txertatuz (*trikitixa, ikastola* eta abar).

5.2 Garapenerako metodologia

Metodologia aldatuz joan da EuskalWordNet-en garapenerako erabilitako irizpideen arabera. Lehenengo urratsak, oinarrizko EuskalWordNet eraikitzea izan zuen xede eta horregatik, estaldura izan genuen helburu nagusi. Hala, garapenaren lehenengo urratsean euskarazko ordainak ingelesezko oinarrizko kontzeptuei (*Base Concepts* izenekoei) eskuz lotu ziren eta ondoren, ingelesezko *synset*-en euskal ordainak hiztegi elebidunak baliatuz (euskara-ingelesa[18]) automatikoki sortu ziren.

Hurrengo urratsean, kalitateari eman zitzaion garrantzi gehiago. Hala ere, kalitatea lantzeko metodologia ezberdinak erabili dira. Hasieran, hizkuntzalariak automatikoki sortu ziren euskarazko *synset* horien kontzeptuz kontzeptuko eskuzko orrazketan murgildu ziren. Hau amaitzean, lotutako hitzen adieren hitzez hitzeko eskuzko orrazketarekin hasi ziren, hau da, hitz batek hiztegian —*Elhuyar Hiztegi Txikia* (Elhuyar, 1998)— zituen adierak EuskalWordNet-en zeudela ziurtatzen eta *synset*-ean dauden ordainak egokiak zirela egiaztatzen saiatzen ziren. Azken finean, aurreko urratseko datu berberak, baina beste ikuspegi osagarri batetik begiratuz.

Orrazketa honen erdibidean ginela, metodologiaz aldatzea erabaki genuen, corpusean arreta gehiago jarriz. Abiapuntu gisa, Fellbaum *et al.*en lana (2001) izan genuen. Autore hauen ustez, hiztegiak hitzen adierak emateaz arduratzen dira, hitz hauen testuinguruetaz gehiegi arduratu gabe. Aldiz, corpusek hitzaren testuinguruari eta erabilerari buruzko informazioa eskaintzen digute. Horrela bada, bi baliabide hauen abantailei probetxu ateratzearren, hiztegiak eta corpusak elkarrekin erabiltzea proposatzen dute.

[18] Morris (1998); Aulestia and White (1990).

Proposamen honetatik abiatuta, EuskalWordNet-en hitzez hitzeko eskuzko orrazketa eta corpus baten etiketatze semantikoa uztartzen hasi ginen. Honela, corpuseko informazioa erabil dezakegu EuskalWordNet orrazteko, garatzeko eta aberasteko. Eta aldi berean, EuskalWordNet-eko *synset*-ak erabilita, eskuz etiketatutako euskarazko corpus semantikoa sortzen ari gara: EuSemcor.

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.

Hurrengo atalean, EuSemcor eta EuskalWordNet-en aldi bereko garapenaren berri ematen dugu.

5.3 Corpus etiketatua: EuSemcor

Lan honen helburua 300.000 hitzeko corpusa euskaraz etiketatzea da EuskalWordNet-eko adierak edo *synset*-ak erabiliaz. Izenak, adjektiboak eta aditzak etiketatu nahi dira. Aldi berean, eta corpusetik lortzen den informazioan oinarrituz, EuskalWordNet-eko *synset*-ak orraztuko dira; hau da, behin 300.000 hitzeko corpusaren etiketatze semantikoa amaituta, EuskalWordNet-ek corpusean agertu diren adiera horiek guztiak izan beharko ditu.

Lan-taldea bost hizkuntzalariez osatua dago: gainbegirale bat, editore bat, bi etiketatzaile eta epaile bat. Editorea EuskalWordNet *editatzen* duena da, hots, *synset*-ak orrazten dituena da. Etiketatzaileek etiketatu beharreko hitzaren agerpenak etiketatzen dituzte (bakoitzak bere aldetik). Azkenik, epaileak bi etiketatzaileen lana erkatuko du eta ezberdin etiketatuta dauden agerpen horiek ebatziko ditu.

Laburki esanda, lan-talde honek jarraitzen duen metodologia hurrengoa da: editoreak aukeratzen ditu landu beharreko hitzak, eta hitz hauen EuskalWordNet-eko *synset*-ak orraztuko ditu. Hitzak orraztu ondoren, editorea, etiketatzaileak eta epailea elkartuko dira hitz horien *synset*-en esanahia ulertzeko. Editoreak, epaileak eta etiketatzaileek hitzen *synset*-ak zeintzuk diren ulertu eta adostu dutenean, etiketatzaileak hitzei dagozkien agerpenak etiketatzen hasiko dira. Etiketatze-lana amaitzean, *synset*-en glosak ere ingelesetik euskarara itzuliko dituzte. Lan hauek guztiak bukatu ondoren, etiketatzaileek editoreari eta epaileari jakinaraziko diete eta etiketatzean izan dituzten gorabeherak azalduko dituzte bilera batean. Gero, epaileak bi etiketatzaileen lana erkatuko du eta ezberdin etiketatuta dauden agerpen horiek ebatziko ditu. Gainera, corpusean hitz horien adiera berririk agertuko balitz, horien berri eman beharko dio editoreari. Azkenik, editoreak corpusean agerturiko adiera berri horien egokitasuna aztertuko du hauek EuskalWordNet-en sartzea erabaki baino lehen.

3. irudian ikus daitekeen bezala, metodologia ziklikoa da.

5.4 Egoera eta erabilera

EuskalWordNet orain dela bost urte garatzen hasi ginen, eta etengabe aberasten doan EBLSa da. Gaur egun, EuskalWordNet-ek 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,

3. irudia: EuSemcor: metodologia.

aurrerago, EuskalWordNet eta *Euskal Hiztegia*ren (Sarasola, 1996) arteko mapaketa edota bateratze bat egitea, hau da, EuskalWordNet *Euskal Hiztegia*tik eratortzen diren erlazio lexiko-semantikoekin aberastea.

EuskalWordNet-en erabilerak era askotakoak izan daitezke. Alde batetik, hiztegi eta thesaurus gisa erabili izan da. Hiztegi tradizionaletan bezala, EuskalWordNet-ek *synset* bakoitzeko glosa bat du, gehienetan adibide eta guzti. 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 (Martínez, 2005).

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, eta 5.3. atalean aipatu bezala, egun, IXA taldean euskarazko corpus bat semantikoki etiketatzean ari gara: EuSemcor (Agirre *et al.*, 2006*a*).

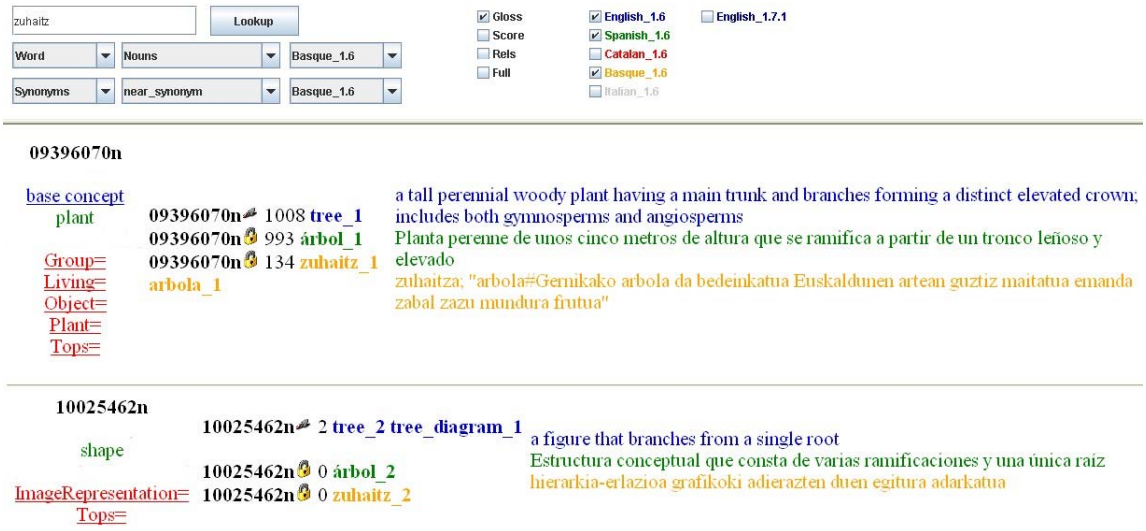
5.5 EuskalWordNet: kontsulta

EuskalWordNet 1.6 bertsioa kontsultagarri dago http://ixa2.si.ehu.es/mcr/wei.html web orrian, eta interfazeren itxura 4. irudikoa da. Interfazearen erabileraren berri eman baino lehen, interfazea ondo erabiltzeko beharrezkoa den oinarrizko terminologia azalduko dugu.

5.5.1 Oinarrizko terminologia

- ***Synset-a:*** Kontsultatu dugun hitzaren adiera ezberdin bakoitzari *synset* bat dagokio, eta interfazean marra batez bereiztuta agertzen da. 4. irudian ikus daitekeen bezala, zuhaitz hitzak bi *synset* ditu, hau da, bi adiera: 'arbola' eta 'diagrama'. Bestalde, *synset* 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. 4. 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.



4. irudia: EuskalWordNet: interfazea.

5.5.2 EuskalWordNet: interfazea

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

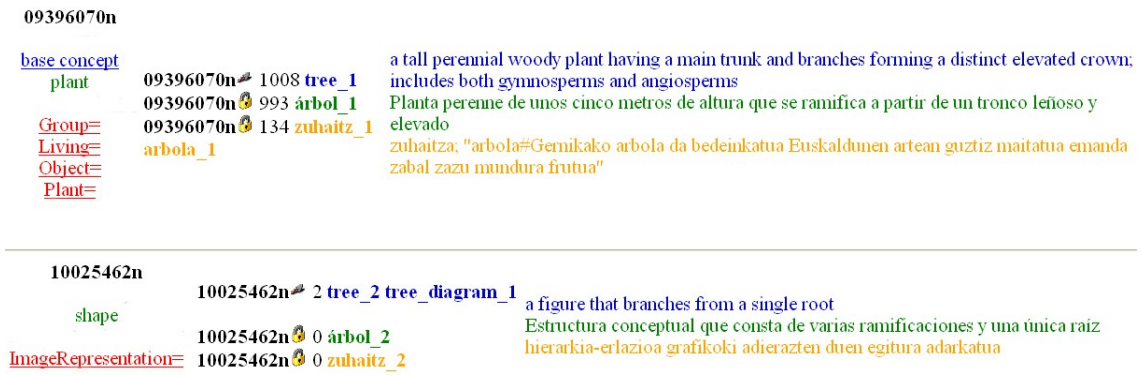
5. irudia: EuskalWordNet: kotsultarako bete beharreko eremuak.

Kontsulta bat egiterakoan, lehenengo testu-kutxatilan kontsultatu nahi den hitza, *synset*-a edo *variant*-a idatzi behar da (ikus 5. irudian A hizkia), eta ondoren, testu-kutxatilan idatzitakoa hitza, *synset* edo *variant*-a den zehaztu behar da testu-kutxatilaren azpiko eremuan (B hizkiaz 5. 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 *Basque1.6* aukeratu ditugu (B hizkiaz 5. 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 5. irudian).

Eta azkenik, hainbat kontrol-laukiei eraginda (D hizkiaz 5. 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.

Kontsulta honen emaitza 6. irudian ikus dezakegu. Alde batetik, *zuhaitz* hitzak bi *synset* dituela adierazten da. Lehenengo *synset*-a 'landare' adierari dagokio, eta bigarrena, berriz, 'diagrama' adierari.



6. irudia: EuskalWordNet: kontsultaren emaitza

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 (ikus 3.3. atala). Marka hau beti berdez adierazita dator. 6. 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 (ikus 3.3. atala). Marka hau beti gorriz adierazita dator. 6. 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 duela eta biziduna 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, 6. 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)[19].

[19]Web orrian interfazeari buruzko informazio gehiago duen eskuliburua dago eskuragarri.

6 ONDORIOAK ETA ETORKIZUNERAKO LANAK

Lan honetan EuskalWordNet ezagutza-base lexiko-semantikoa aurkezteaz gain, EBLS hau lortzeko jarraitutako prozedura ere deskribatu dugu: euskararako baliagarria izan zitekeen eredu baten bila hasi ginenetik, aukeratutako eredua euskararako garatu eta aplikatu dugun arte. Horretarako, lehenengo euskararako erabilgarria izan zitekeen EBLS baten proposamena egiteko zer eredu eta irizpidetan oinarritu garen azaldu dugu, eta ondoren, WordNet eta honen ildotik abiatutakoen eredua aukeratu izanaren arrazoiak azaldu ditugu.

EuskalWordNet-en garapenari buruz ere jardun gara, eta esan beharra dago, EuskalWordNet garapenean dagoela oraindik. Izenak nahiko landuta daude, eta azken orrazketa egiten ari gara horien gainean. Aditzei dagokionez, oinarrizko kontzeptuak (*Base Concept*-ak) daude landuta bakarrik, eta orain masiboki nola landu aztertzen ari gara. Adjektiboen kasuan, oso gutxi daude landuta, eta aurreragorako utzi ditugu.

EuskalWordNet lantzen ari garen heinean, euskararako aukeratutako EBLS ereduak —WordNet eta honen ildotik abiatutakoak, alegia— ez digu arazorik eman. Hala ere, eredua aukeratzean, honen gabezia ezagutzen genuen, informazio sintaktiko-semantikoa mugatua duela, hain zuzen ere. Arazo honi aurre egiteko, euskarako EBLSa MCRren ereduan eraikitzea erabaki dugu, informazio sintaktiko-semantiko hori eransteko erraztasuna baduela ikusi dugulako. Hala, honi esker, aztertu ditugun gainontzeko beste ereduetatik (Levin (1993), FrameNet, eta abar), gure euskarako EBLSrako baliagarria zaigun informazio sintaktiko-semantikoarekin aberasteko aukera dugu. Honenbestez, EuskalWordNet informazio sintaktiko-semantikoarekin osatzen hasiak gara (Pociello, 2004*b*).

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 tematikoak desanbiguatu eta esleitu beharko dira, eta horretarako rol tematikoz etiketatutako corpus bat garatzen ari gara, EusPropBank deritzona (Civit *et al.*, 2005).

Bestalde, interpretazio semantikorako beharrezkoak diren hainbat eta hainbat erlazio lexiko-semantiko txertatu nahi dira (adibidez garagardoa garagarrez egina dagoela, okinaren lanbidea ogia egitea dela, arrantza pertsonek —eta astoek— egiten dutela, etab.). Aberasketa horretarako *Euskal Hiztegia* eta corpusetatik ezagutza hori erdiautomatikoki erauzten saiatzen diren teknikak ere lantzen ari gara (Lersundi, 2005; Ansa *et al.*, 2005).

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

Aplikazioei dagokionean, EuskalWordNet itzulpen automatikoan eta informazio bilaketan erabiltzeko planak dauzkagu. Gure taldean garatu den informazio bilatzaile eleanitz batean erabili dugu jada, eta oso emaitza onak lortu ditugu (Vossen *et al.*, 2006). Sistema honetan erabiltzaileak hainbat hizkuntzatan egin dezake bilaketa (euskara, katalana, italiera, gaztelania eta ingelesa), eta emaitza EFE enpresako gaztelania eta ingeleseko oina duten argazkiak osatzen dute. Sistema honek azaltzen ditu hobekien WordNet eredua hartzearen abantailetako batzuk: beste hizkuntzetarako loturak "dohainik" lortzea, alde batetik, eta eredu horretan oinarritutako sistemetan hizkuntza berri bat gehitzeko

erraztasuna.

Gogora dezagun, EuskalWordNet helbide honetan atzitu daitekeela ingelesezko interfaze baten bidez: http://ixa2.si.ehu.es/mcr/wei.html. Bestalde, EuskalWordNet-i buruzko argibide gehiago Agirre *et al.* (2005, 2006, *b*) lanetan aurki daiteke.

Erreferentziak

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

Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, E., Mendizabal, K., Quintian, M. and Pociello, E. (2005) EuskalWordNet: euskararako ezagutza-base lexiko-semantikoa. *Euskalingua* (7).

Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, E., Mendizabal, K., Quintian, M. and Pociello, E. (2006*a*) Improving the Basque WordNet by corpus annotation. In *Proceedings of Third International WordNet Conference*, Jeju, Korea.

Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, E., Mendizabal, K., Quintian, M. and Pociello, E. (2006*b*) A methodology for the joint development of the Basque Wordnet and Semcor. In *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC)*, Genoa, Italy.

Agirre, E., Aldezabal, I. and Pociello, E. (2003) A pilot study of English selectional preferences and their cross-lingual compatibility with Basque. In *Proceedings on International Conference on Text Speech and Dialogue (TSD)*, Czech Republic.

Agirre, E., Aldezabal, I. and Pociello, E. (2006) Lexicalization and multiword expressions in the Basque Wordnet. In *Proceedings of Third International WordNet Conference*, Jeju, Korea.

Agirre, E., Alegria, I., Arregi, X., Artola, X., de Ilarraza, A.D., Maritxalar, M., Sarasola, K. and Urkia, M. (1992) Xuxen: a spelling checker/corrector for Basque based in two-level morphology. In *Proceedings of ANLP'92*, Povo Trento.

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

Agirre, E. and Martínez, D. (2000) Exploring automatic word sense disambiguation with decision lists and the Web. In *Proceedings of the Semantic Annotation And Intelligent Annotation workshop organized by COLING*, Luxembourg.

Agirre, E. and Martínez, D. (2001) Learning class-to-class selectional preferences. In *Proceedings of the Workshop "Computational Natural Language Learning"*, Toulouse, France.

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

Aldezabal, I., Aranzabe, M., Atutxa, A., Gojenola, K., Sarasola, K. and Goenaga, P. (2001) Extracción masiva de información sobre subcategorización verbal vasca a partir de corpus. In *Actas del XVII Congreso de la SEPLN Universidade de Jaén*, Jaén, Spain.

Ansa, O., Arregi, X., Esparza, I. and Valverde, A. (2005) Un entorno para el desarrollo y la evaluación de un sistema de búsqueda de respuestas en euskera. In *Proceedings of the Annual SEPLN meeting*, Granada, Spain.

Arriola, J. (2000). *Euskal Hiztegia-ren azterketa eta egituratzea ezagutza lexikalaren eskuratze automatikoari begira. Aditz-adibideen analisia Murriztapen-gramatika baliatuz, azpikategorizazioaren bidean*. PhD thesis, UPV-EHU.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B. and Vossen, P. (2004) The MEANING Multilingual Central Repository. In *Proceedings of the 2nd Global WordNet Conference*, Brno, Czech Republic.

Aulestia, G. and White, L. (1990) *English-Basque Dictionary*. University of Nevada Press.

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

Buitelaar, P. (1998). *Systematic Polysemy and Underspecification*. PhD thesis, Brandeis University.

Carroll, J., Rigau, G., Magnini, B., Agirre, E., Rodrígez, H. and Atserias, J. (2003). MEANING: cycle 1: acquisition. Technical report.

Civit, M., Aldezabal, I., Pociello, E., Taulé, M., Aparicio, J., Màrquez, L., Navarro, B., Castellví, J. and Martí, M. (2005) 3LB-LEX: léxico verbal con frames sintácticos-semánticoss. In *Revista de la Asociación Española para el Procesamiento del Lenguaje Natural*, Granada, Spain.

Dorr, B. (1993) Machine translation. A view from the lexicon. *Computational Linguistics,* 20 (4).

Dorr, B. (1997) Large-scale acquisition of LCS-based lexicons for foreign language tutoring. Washington, DC.

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

Elhuyar (1998) *Elhuyar Hiztegi Txikia*. Elhuyar Kultur Elkartea.

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

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

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

Fillmore, C.J. (1985) Frames and the semantics of understanding. In *Quaderni di Semantica* vol. 6.2,.

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

Gojenola, K. (1998) *Guneak zuzendutako egitura sintagmatikoen gramatika (HPSG) eta Euskararako aplikazioa*.

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

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

Lersundi, M. (2005). *Ezagutza-base lexikala eraikitzeko Euskal Hiztegiko definizioen azterketa sintaktiko-semantikoa. Hitzen arteko erlazio lexiko-semantikoak: definizio-patroiak, eratorpena eta postposizioak*. PhD thesis, UPV-EHU.

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

Lyons, J. (1977) *Semantics*. Cambridge University Press.

Magnini, B. and Strapparava, C. (2001) Using WordNet to improve user modelling in a web document recommender system. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh.

Mandala, R., Takenobu, T. and Hozumi, T. (1998) The use of WordNet in information retrieval. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal.

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

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

Miller, G.A., Chodorow, M., Landes, S., Leacock, C. and Thomas, R.G. (1994) Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*, San Francisco.

Morris, M. (1998) *Morris Student*. Klaudio Harluxet Fundazioa.

Pociello, E. (2004*a*). Sintaxi-semantika elkargunea zenbait teoriatan: euskararen ezagutza-basea lexiko-semantikorantz. Master's thesis UPV-EHU.

Pociello, E. (2004*b*). Aditzen hautapen-murriztapenak: kirol domeinura mugatutako ingeleseko hautapen-murriztapenak eta euren baliagarritasuna euskararako. Hastapeneko lana. Master's thesis UPV-EHU.

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

Rigau, G., Agirre, E. and Atserias, J. (2003) The MEANING project. In *Proceedings of the "XIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, Alcalá de Henares (Madrid).

Rigau, G., Rodríguez, H. and Turmo, J. (1995) Automatically extracting translation links using a wide coverage semantic taxonomy. In *Proceedings of the 15th International Conference in Language Engineering, IA-95*, Montpelier, France.

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

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

Vossen, P. (1997) EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, Zurich.

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

Vossen, P., Rigau, G., Alegria, I., Agirre, E., Farwell, D. and Fuentes, M. (2006) Meaningful results for Information Retrieval in the MEANING project. In *Proceedings of Third International WordNet Conference*, Jeju Island, Korea.